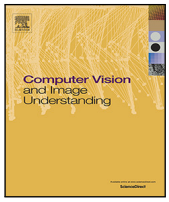
Computer Vision and Image Understanding 243 (2024) 103978

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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



Visual tracking in camera-switching outdoor sport videos: Benchmark and baselines for skiing

Matteo Dunnhofer *, Christian Micheloni

Machine Learning and Perception Lab, University of Udine, via delle Scienze 206, 33100 Udine, Italy

A B S T R A C T

Skiing is a globally popular winter sport discipline with a rich history of competitive events. This domain offers ample opportunities for the application of computer vision to enhance the understanding of athletes' performances. However, this potential has remained relatively untapped in comparison to other sports, primarily due to the limited availability of dedicated research studies and datasets. The present paper takes a significant stride towards bridging these gaps. It conducts a comprehensive examination of skier appearance tracking in videos capturing their entire performance—an essential step for more advanced performance analyses. To implement this investigation, we introduce SkiTB, the largest and most annotated dataset tailored for computer vision applications in skiing. We subject a range of visual object tracking algorithms to rigorous testing, including both well-established methodologies and a novel skier-specific baseline algorithm. The results yield valuable insights into the suitability of various tracking techniques for vision-based skiing analysis and into the generalization of state-of-the-art algorithms to complex target behaviors and conditions set by winter outdoor environments. To foster further development, we make SkiTB, the associated code, and the obtained results accessible through https://machinelearning.uniud.it/datasets/skitb.

1. Introduction

Skiing, a recreational winter human activity of global renown (Vanat, 2022), has a rich historical legacy of competitive events dating as far back as the 1840s. Since its inception in 1924, this sport discipline has occupied a central position in the spectacularity of the Winter Olympic Games, as documented by the (International Olympic Committee, 2023). In the contemporary era, skiing encompasses a diverse array of disciplines, including alpine skiing, ski jumping, and freestyle skiing, as recognized by the International Ski and Snowboard Federation (FIS). These various skiing disciplines collectively hold a prominent status within the winter sports industry, garnering substantial attention with over 1.7 billion media views during a typical winter season, according to Nielsen reports (The Nielsen Company, 2022a,c,b).

Leveraging data-driven analytics in the context of skiing performance has the potential to: enhance athletes' technical skills; promote their physical well-being; enrich the educational content and elevate the entertainment factor of broadcasting professional competitions. These advancements contribute to creating more remarkable, secure, and captivating sport competitions. In such applications, computer vision offers promising opportunities for capturing and analyzing skiing performances without relying on wearable sensors, as it has been demonstrated for other sport disciplines (Thomas et al., 2017). Indeed, the effort of applying and researching computer vision methods in skiing has been limited compared to other sports such as soccer (Vandeghen et al., 2022; Honda et al., 2022; Cioppa et al., 2022; Theiner et al., 2022; Gadde and Jawahar, 2022; Theiner and Ewerth, 2023), basketball (Bettadapura et al., 2016; Bertasius et al., 2017; Li and Chuah, 2018; Quiroga et al., 2020; Chappa et al., 2023), or ice-hockey (Pidaparthy and Elder, 2019; Koshkina et al., 2021; Vats et al., 2021; Pidaparthy et al., 2021; Vats et al., 2022).

Previous research in the skiing domain has primarily focused on reconstructing skiers' poses in 2D or 3D (Ludwig et al., 2022; Bachmann et al., 2019) and understanding the style of ski jumps (Štepec and Skočaj, 2022; Wang et al., 2019). A crucial step in building these vision-based analytical tools involves localizing the skier appearance within the video frames. This is an essential computer perception task whose output influences the accuracy of the subsequent higher-level computational modules. The methods usually rely on off-the-shelf or fine-tuned object detection models (Ren et al., 2015; Redmon et al., 2016; Liu et al., 2016) without utilizing the temporal information available in the athlete's performance evolution captured in video. Additionally, the limited and sparsely labeled datasets used in previous studies represent a significant obstacle to the widespread development and applicability of computer vision algorithms in skiing. Skiing videos constitute a particular setting of the task of high-level human activity understanding which presents several unique challenges characterized by exercises performed with unique body-equipment relations, at high speed, and on a continuously changing and widely extended playing field subject to extreme outdoor winter weather conditions (Cheng et al., 2023). All of these circumstances raise important questions regarding their influence on image and video-based systems. Addressing them in a systematic and extensive manner could have implications not only

* Corresponding author.
E-mail address: matteo.dunnhofer@uniud.it (M. Dunnhofer).

https://doi.org/10.1016/j.cviu.2024.103978
Received 21 September 2023; Received in revised form 30 December 2023; Accepted 26 February 2024
Available online 19 March 2024

for the communities orbiting around skiing but even for the computer vision community as a whole as evidenced by recent activities.[1]

For these motivations, this manuscript presents an extensive study on the fundamental task of tracking an athlete (i.e., a skier) appearance in monocular broadcasting videos of professional skiing competitions. Due to the unavailability of suitability benchmarks, a new visual tracking dataset named SkiTB ("Skiers from the Top to the Bottom") is introduced to implement the investigation. SkiTB consists of 300 video recordings across the most challenging skiing disciplines. The videos cover the athletes' complete performance, from the top to the bottom of the course,[2] as exemplified by Fig. 1. Considering the large spatial extent of courses on mountain slopes, multiple cameras are placed in sequential order along the slope to capture the complete skiing performance in such videos. Each video is densely labeled with the bounding-boxes of a single target skier and with attributes identifying the camera ID, the visual changes that the skier undergoes (e.g., different color appearance due to changing illumination conditions, small scale because of camera zoom settings, partial visibility due to occluding items, or perturbed shape due to motion blur), the type of skiing discipline, the athlete ID, the location of the competition, the weather conditions, as well as the parameters of the skiing performance. SkiTB offers multiple training and test splits, making it suitable for developing learning-based computer vision algorithms. We use this benchmark to extensively evaluate different tracking algorithms, including established methodologies and a newly introduced skier-optimized baseline algorithm. Standard protocols and metrics are adapted and utilized to evaluate the specific challenges of video tracking in skiing. The impact of these tracking algorithms on higher-level skiing performance understanding tasks is also investigated. The results provide valuable insights into the applicability of different visual tracking methods for skiing analysis, and the robustness and generalization of state-of-the-art algorithms to challenging factors represented by the domain.

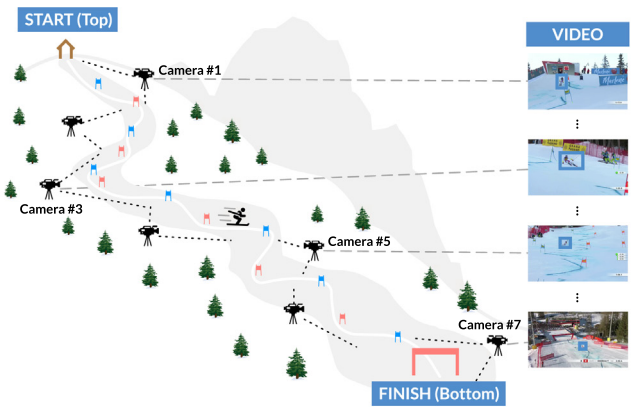
In short, the contributions of this paper are:

- A systematic and in-depth investigation of the problem of skier tracking in videos compiled from clips acquired by multiple cameras (i.e., videos with camera shot-cuts), which has not been thoroughly studied in previous works.
- The description and release of SkiTB a novel benchmark dataset curated specifically for evaluating and developing computer vision-based systems in the skiing domain. The dataset is designed to be diverse, representative, and densely labeled.
- STARK$_{\texttt{SKI}}$, a baseline algorithm optimized for skier tracking in videos characterized by camera switching operations.

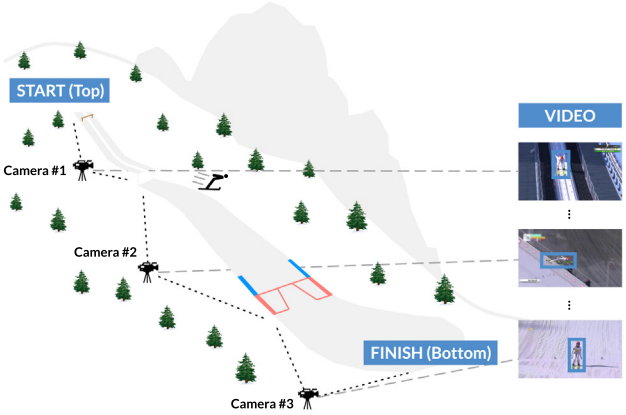
2. Related work

2.1. Visual object tracking

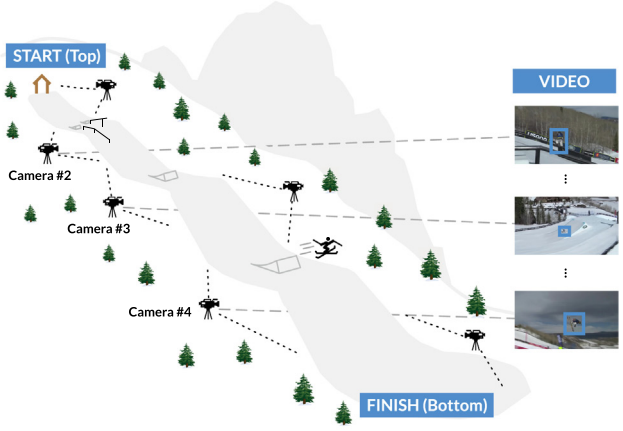
In the recent past, there has been increasing interest in developing precise and robust single object tracking (SOT) algorithms for various domains (Chen et al., 2022). Early trackers utilized mean shift algorithms (Comaniciu et al., 2000), key-point (Maresca and Petrosino, 2013), part-based techniques (Čehovin et al., 2013), or SVM learning (Hare et al., 2016). Correlation filters gained popularity due to their fast processing (Bolme et al., 2010; Henriques et al., 2015). More recently, deep learning-based solutions, including regression networks (Held et al., 2016), online tracking-by-detection methods (Nam and Han, 2016), reinforcement learning-based methods (Yun et al., 2017; Choi et al., 2018; Dunnhofer et al., 2019, 2020), deep discriminative correlation filters (Danelljan et al., 2019; Bhat et al., 2019), siamese network-based trackers (Bertinetto et al., 2016; Li et al., 2019;



(a) Alpine Skiing - AL



(b) Ski Jumping - JP



(c) Freestyle Skiing - FS

Fig. 1. Tracking a skier from the top to the bottom of the course. This paper focuses on applying visual object tracking algorithms to localize a skier per-frame (e.g. with bounding-boxes □) in a video capturing his/her complete performance. Due to the large spatial extent of skiing courses, multiple cameras (typically pan–tilt-zoom) are placed sequentially along the slope to capture the whole performance, and multi-camera tracking is required for high-level performance analysis. This figure shows such a camera setup for the skiing disciplines of alpine skiing (a), ski jumping (b), and freestyle skiing (c).

[1] 1st Workshop on Computer Vision for Winter Sports at WACV 2022 https://machinelearning.uniud.it/events/CV4WS-2022
2nd Workshop on Computer Vision for Winter Sports at WACV 2023 https://machinelearning.uniud.it/events/CV4WS-2023

[2] In the scope of this paper, a skiing course is considered as a path or track down a mountain slope that an athlete should follow to complete his/her performance. It should not be confused with a course taken to learn how to ski.

Table 1
Comparison of SkiTB with publicly available skiing-related datasets. This table shows a comparison between some key statistics of our SkiTB and the other datasets for computer vision tasks available in the skiing domain: Skimovie (Steinkellner and Schöffmann, 2021), Ski2DPose (Bachmann et al., 2019), SkiPosePTZ (Bachmann et al., 2019), YouTube Skijump (Ludwig et al., 2023). As can be noticed, ours results in the largest, most diverse, and most annotated dataset. (n/a stands for "not annotated".)

Dataset	Skimovie	Ski2DPose	SkiPosePTZ	YouTube Skijump	SkiTB
Skiing application	Detection	2D Pose Estimation	3D Pose Estimation	2D Pose Estimation	Tracking
Per-frame annotations	✓ (12.5 FPS)	n/a	n/a	n/a	✓ (30 FPS)
Complete performance	✓	n/a	n/a	n/a	✓
Performance parameters	n/a	n/a	n/a	n/a	✓
Weather annotations	n/a	n/a	n/a	n/a	✓
# multi-camera videos	n/a	n/a	n/a	n/a	300
# single-camera videos	4	n/a	36	n/a	2019
# annotated frames	2718	1982	20K	2867	352 978
# skiing disciplines	1 (AL)	1 (AL)	1 (AL)	1 (JP)	3 (AL, JP, FS)
# sub-disciplines	1	4	1	2	11
# athletes	n/a	32	6	118	196
# locations	6	5	1	7	161

Lu et al., 2023), and transformers (Yan et al., 2021; Cui et al., 2022; Ye et al., 2022; Mayer et al., 2022; Lin et al., 2022), led to higher tracking accuracy. Long-term trackers and methods combining multiple trackers have been also explored (Yan et al., 2019; Huang et al., 2020; Dunnhofer and Micheloni, 2022).

Such a progress in SOT algorithms is attributed to well-curated evaluation datasets featuring diverse object types (Wu et al., 2015; Kristan et al., 2020, 2021, 2023) and large-scale datasets for visual object tracking in generic domains (Müller et al., 2018; Huang et al., 2019; Fan et al., 2019). Application-centric benchmarks exist for specific domains such as drones (Mueller et al., 2016), high frame-rate videos (Galoogahi et al., 2017), transparent objects (Fan et al., 2021), and egocentric videos (Dunnhofer et al., 2023a). These benchmarks contribute to the development of accurate and reliable tracking systems in specific application scenarios.

The aforementioned datasets (Wu et al., 2015; Kristan et al., 2023; Huang et al., 2019; Müller et al., 2018; Fan et al., 2019) lack a sufficient representation of skiing, hindering the development of effective trackers in this domain. To overcome this limitation, we introduce SkiTB as a comprehensive and well-curated benchmark for evaluating trackers on skiing regardless of their methodology. The dataset covers the unique aspects of the skiing domain, including fast human motion, extreme weather conditions, and distractor objects. We believe that SkiTB can also benefit the development of generic tracking methodologies.

2.2. Visual tracking in sport videos

Our investigation builds upon prior research in the realm of athlete tracking within video footage. Mauthner et al. (2007) introduced an algorithm based on integral histograms for the tracking of volleyball players in video sequences obtained from a single camera source. Kristan et al. (2009) leveraged closed-world assumptions with respect to both visual and dynamical characteristics of players engaged in indoor sports such as handball and basketball, as captured from a bird's-eye view camera perspective. Liu et al. (2013) proposed the utilization of context-conditioned motion models that implicitly incorporate intricate inter-object correlations to facilitate the tracking of multiple athletes involved in basketball and field hockey. Morimitsu et al. (2017) explored the application of structural relations between the athletes' positions for tracking them in single-camera video footage encompassing table tennis, badminton, and volleyball. Cui et al. (2023) have made available an extensive and densely-labeled dataset comprising video recordings of basketball, volleyball, and football. This dataset has served as a benchmark for evaluating state-of-the-art multiple object tracking (MOT) algorithms (Dendorfer et al., 2021). A parallel investigation focusing solely on soccer was conducted by Cioppa et al. (2022). The latter built open the study performed by Feng et al. (2020) on single player tracking in the same sport discipline.

In contrast to the aforementioned studies, our investigation uniquely centers on the sport of skiing. Unlike team sports, skiing represents a discipline wherein athletes strive to optimize their individual performance. Consequently, the development of an automated and effective video analysis system necessitates the inclusion of a visual perception system featuring a SOT algorithm and not an MOT one. Furthermore, in contrast to single-camera video stream used in the previous works, our investigation is centered on multi-camera videos with camera shot-cuts. This is set by the specific characteristics of the field on which skiing is performed, which necessitates the generation of a video sequence captured by several cameras to observe and analyze the complete performance. In this view, the study most closely related to ours is the one conducted by Drory et al. (2017), which introduced a visual tracking algorithm designed for tracking kayak athletes in videos that contain shot-cuts. Nevertheless, our work is distinctive in its exclusive focus on skiing, an outdoor discipline characterized by distinct visual attributes, including rapid motion and substantial variability arising from intra-discipline differences and exposure to variable weather conditions (Cheng et al., 2023).

2.3. Applications of computer vision to skiing

Recent advancements in computer vision (He et al., 2016; Ren et al., 2015; Cao et al., 2017) have enabled vision-based applications in skiing performance analysis. For example, Zhu and Yan (2022) proposed object detection and human pose estimation algorithms to recognize falls of alpine skiers, while Zwölfer et al. (2021) discussed the combination of pose estimation with kinematics models. Bachmann et al. (2019) introduced a methodology to reconstruct 3D poses from images captured by multiple synchronized cameras observing a single slope section. Ski jumping analysis involved scoring the style of jumps using 2D human pose trajectories (Štepec and Skočaj, 2022) and detecting key-points on the human body and skis in still images using improved vision transformer architectures (Ludwig et al., 2022, 2023). For freestyle skiing and snowboarding, algorithms were developed to evaluate the quality of jumps in monocular videos (Wang et al., 2019) and to synchronize videos for comparing the timing and spatial extent of aerial maneuvers (Matsumura et al., 2021).

The discussed pipelines present object detection (Ren et al., 2015; Redmon et al., 2016) or off-the-shelf visual tracking (Wang et al., 2019) for initial skier localization, followed by subsequent modules for higher-level output computation. The accuracy of skier localization greatly affects the performance of the successive modules, but this aspect has been overlooked by existing systems. Only limited evaluations on skier localization accuracy have been conducted in previous works (Steinkellner and Schöffmann, 2021; Qi et al., 2022). These studies focused on a small number of videos and lacked analysis of the challenging characteristics of the skiing domain. In contrast, this paper presents a systematic and comprehensive analysis of skier tracking on a large scale, involving 300 videos and 353K frames. Multi-camera-switching videos capturing professional athletes from various skiing disciplines were used, considering real competition conditions

with different courses, skiing styles, distracting skiers, and challenging weather. A comparison between the proposed SkiTB and publicly available computer vision datasets for skiing applications is presented in Table 1.

3. Problem formulation

This paper focuses on the per-frame localization of a specific skier in a video stream capturing his/her complete performance on a skiing course, from the top of such a skiing course to its bottom. The video stream is a sequence $\mathcal{V} = \{F_t \in \mathcal{I}\}_{t=0}^{T}$ of frames F_t, where $\mathcal{I}$ represents the space of RGB images and T is the total number of frames. The bounding-box $b_t = [x_t, y_t, w_t, h_t] \subseteq \mathbb{R}^4$ defines the position and size of the skier's appearance in each frame, and the objective is to develop a visual tracking algorithm – also referred to as tracker – to predict the bounding-box b_t with a confidence value $0 \leq c_t \leq 1$, for $0 < t \leq T$, in an online fashion. The initial bounding box b_0 can be generated by an object detection algorithm (Redmon et al., 2016; Ren et al., 2015) or manually annotated by a human operator. Skiing competitions involve courses spanning several hundred meters if not kilometers, requiring multiple cameras to be placed sequentially along the slope to capture the skier's entire performance. Thus, $\mathcal{V}$ consists of frames grabbed by several different cameras and concatenated into a single stream showing a complete performance. Considering that skiing is an individual sport, our problem of interest constitutes an application case of single long-term object tracking (Lukežič et al., 2020; Kristan et al., 2023), specifically of the global instance variation (Hu et al., 2023) which aims to continuously localize a target object over an extended period, even across camera shot-cuts. Fig. 1 presents a visualization of such a setting for the case of alpine skiing. We assume that manual camera control and camera switching occur, as it is done for real-time broadcasting transmission. Our paper focuses on the problem of per-frame skier appearance localization in a single video. It should not be confounded with the problem of multi-camera target tracking (Ristani et al., 2016; Rhodin et al., 2018) where targets are located by exploiting multiple videos acquired by several synchronized and mutually calibrated cameras. We believe that the findings resulting by the study of our problem of interest can contribute to the development of such technologies that could be capable of tracking a skier in the 3D space. Indeed, to exploit multi-view geometry algorithms across views (Hartley and Zisserman, 2003) and to eventually control automatically such cameras, the athlete's appearance must be first localized in the frames of the single video stream captured by each cameras.

4. The SkiTB dataset

The SkiTB dataset provides a comprehensive spatio-temporal video representation and annotation of professional skiing performance under the settings described in Section 3. SkiTB comes with dense annotations for tracking purposes, but it is designed to serve as a well-curated benchmark for subsequent higher-level skiing performance understanding tasks. In particular, we adhere to the following design principles:

- Scale: we ensured that SkiTB would contain a significant number of videos and frames to facilitate the development of modern computer vision solutions based on deep learning.
- Diversity: we included a wide range of situations, such as different skiing disciplines, athletes, skiing styles, courses and locations, to enable the testing and generalization of methods under various application conditions.
- Representativeness: we designed SkiTB to represent real competition scenarios of professional athletes, which enables the development of algorithms capable of working in real-world situations.

Table 2
Key statistics of SkiTB. The following table offers overall and per-skiing discipline (AL: alpine skiing, JP: ski jumping, FS: freestyle skiing) information about the multi-camera (MC) and single-camera (SC) videos and the associated data present in the proposed dataset.

Skiing discipline	AL	JP	FS	All
# MC videos	100	100	100	300
# SC videos	1100	346	573	2019
# frames	215 517	38 201	99 260	352 978
# cameras (min, avg, max)	(6, 11, 26)	(2, 3, 5)	(1, 6, 15)	(1, 7, 26)
avg MC video seconds	71	13	33	39
avg SC video seconds	6.5	3.6	5.7	5.8
# sub-disciplines	4	2	5	11
# athletes	56	54	86	196
# athlete genders (M, W)	(34, 22)	(35, 19)	(49, 37)	(118, 78)
# athlete nationalities	15	10	18	25
# courses	68	34	59	161
# courses countries	15	12	17	24

Fig. 2. Frame and bounding-box samples from SkiTB. We showcase examples of video frames from our dataset for the different disciplines: alpine skiing (AL), ski jumping (JP), and freestyle skiing (FS). Each frame is accompanied by a manually annotated bounding-box. A blue rectangle (□) localizes the skier's appearance as visible, while a black rectangle (□) as occluded. The camera that captured the frame and the elapsed time in seconds from the beginning of the performance are also reported. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The challenge represented by SkiTB arises from the complex and dynamic nature of skiing and its environment, where an athlete's visual appearance and motion are influenced by highly variable factors such as: complex body movements due to high speed, course settings, aerial execution, impact absorption; particular image characteristics due to meteorological conditions (e.g., snowing, raining, and intense shadowing) and camera operations (e.g., camera switching, fast camera movements, long-range capturing). From a more general point of view, SkiTB can serve as a valuable resource for research in multi-camera target tracking under extremely dynamic outdoor environments.

4.1. Video collection

SkiTB contains 300 videos carefully selected from broadcasting recordings showcasing complete skiing performances available on the Internet. Our selection process aimed to maximize diversity in terms

of athletes, locations, courses, and weather conditions, while ensuring a balanced distribution across three major skiing disciplines defined by the FIS rules (International Ski and Snowboard Federation, 2023): alpine skiing (AL), ski jumping (JP), and freestyle skiing (FS). These disciplines were chosen based on their popularity and the challenging representations they provide in videos. Existing datasets (Steinkellner and Schöffmann, 2021; Bachmann et al., 2019; Ludwig et al., 2023) did not encompass all the desired characteristics, necessitating the creation of a new video collection. The videos have a framerate of 25 or 30 FPS and resolutions ranging from 360p to 720p. More detailed statistics can be found in Table 2.

4.2. Frame-level annotations

Each of the 300 videos is composed as a single stream of frames acquired by multiple different cameras. We refer to each of these videos as a multi-camera (MC) video. Each of the frames belonging to the 300 MC videos has been manually labeled with the bounding-box enclosing the visual appearance of the athlete and its equipment (skis, and poles if present), as shown in Fig. 2. The sequence of boxes for each video starts with a frame capturing at least 50% of the skier's appearance shortly before the descent begins, and it ends with a frame capturing the skier after completing their performance. Each box is labeled to indicate whether the skier is visible or occluded (*i.e.*, when approximately more than 50% of the skier's visual appearance is hidden). The dataset includes instances of complete occlusions, such as when the skier passes behind snow ramps in FS. In such cases, boxes are drawn to localize the skier in likely positions based on the observed motion. On average, complete occlusions last for 15 frames. The motivation behind the employment of bounding-boxes is grounded on the fact that such a representation is sufficiently informative for the computational processes performed by higher-level skiing performance understanding tasks (Bachmann et al., 2019; Ludwig et al., 2022, 2023; Štepec and Skočaj, 2022). The aforementioned pipelines simply require a rectangle highlighting the area covered by the skier's appearance. Compared to the more complex segmentation masks (Kristan et al., 2020, 2021), the four-value representation of bounding-boxes demands less computational resources, thus enables the development of more efficient methods. Additionally, the choice of including the appearance of the skiing equipment within the labeled bounding-box is guided by the common working mechanism of the aforementioned solutions, which necessitate a bounding-box encompassing both the athlete's body and equipment.

Each frame is also labeled with the index of the camera that captured it. The camera order for each video was manually determined by assessing the order of video shot-cuts. This enumeration reflects the sequence in which the cameras were positioned along the slope.

Some video frames include virtual graphics showing the performance results of the athlete. Based on experiments, we have seen that the algorithms' behavior is not significantly influenced by such a presence, thus we treated the graphics as objects occluding the captured scenes.

4.3. Video-level and clip-level annotations

To enable in-depth analysis, we have associated labels with both the MC videos and the single-camera (SC) clips, which are sub-sequences of frames captured by the same camera. Each MC video is labeled with the following information: the discipline (AL, JP, FS); the specific skiing sub-discipline; the visible weather condition; athlete ID (including name and surname) and nationality; the date, location, and country of the competition. It is worth noting that each MC video is also annotated with the athlete's performance results in computable form, even though these labels are not specifically utilized in this work. Fig. 3 shows that the representation of athlete nationalities and course locations

Fig. 3. World-wide data representation. The diversity of the videos present in SkiTB is demonstrated by the nationalities of the athletes and the skiing courses represented in our dataset. SkiTB features almost all the countries in which winter sports are very popular activities.

associated to the MC videos covers all the major countries in which winter sports are majorly popular.

Each SC clip is associated with labels (CM, SC, BC, ARC, IV, POC, MB, FM, FOC, LR) that express the visual variability of the target skier, reinterpreted to suit the application domain. Table 3 presents the description and the application-wise interpretation of such attributes. Fig. 4 shows the distribution of the SC clips according to the labels. All of these labels can be utilized for video clustering, enabling experimental results to be conditioned on different characteristics of the domain. This evaluation approach is well-established in the visual object tracking community (Wu et al., 2015; Galoogahi et al., 2017; Dunnhofer et al., 2023a; Mueller et al., 2016; Fan et al., 2019; Huang et al., 2019) and was shown to be sufficiently robust to estimate the trackers' performance in particular scenarios. Among the many attributes present in the literature, we selected 10 that well represent the variability of the skiing domain. The labels have been associated with SC clips because the SC experimentation setting allows a tracker to cover the situations happening during the skier's descent in a more complete and consistent way (Kristan et al., 2020). Indeed, considering a long video as defined by the MC setting, it could be the case that a full occlusion (FOC) happens in the section captured by Camera #2 causing the tracker to fail. Later in the descent in the section captured by Camera #6,

Table 3
Selected sequence attributes associated to single-camera (SC) clips. This table gives the formal definition of the selected clip attributes according to previous research in generic visual object tracking (Wu et al., 2015; Fan et al., 2019; Dunnhofer et al., 2023a). On a side, we give an interpretation of each definition with respect to our application domain.

Attribute	Definition	Application-specific Interpretation
CM	Camera Motion: an abrupt camera motion can be seen in the video clip.	The camera operator moves the camera fast to keep the skier in the field of view.
SC	Scale Change: the ratio of the bounding-box area of the first and the current frame is outside the range [0.5, 2].	The size of a skier's appearance changes considerably during the video (e.g. by zooming in/out on the target).
BC	Background Clutter: the target has a similar appearance w.r.t. the surrounding background.	The appearance of the athlete's suit and equipment confounds with the elements in the background.
ARC	Aspect Ratio Change: the ratio of the bounding-box aspect ratio of the first and the current frame is outside the range [0.5, 2].	The ratio between the height and width of the athlete changes (e.g. due to complex body poses).
IV	Illumination Variation: the area of the target bounding-box is subject to light variation.	The appearance of the target skier changes due to particular lightning conditions (e.g. passing through slope areas under shadow).
POC	Partial Occlusion: the target is partially occluded in the video.	Part of the skier is hidden by another item (e.g. by a gate in AL).
MB	Motion Blur: the target region is blurred due to target or camera motion.	The appearance of the skier is blurred due to its fast motion or the fast motion of the camera.
FM	Fast Motion: the target bounding-box has a motion change larger than its size.	The skier moves fast during the descent on the course.
FOC	Full Occlusion: the target is fully occluded in the video.	The skier is completely occluded by another item in the field of view (e.g. by a kicker in FS).
LR	Low Resolution: the area of the target bounding-box is less than 1000 pixels in at least one frame.	The skier appears small due to a low level of camera zoom.

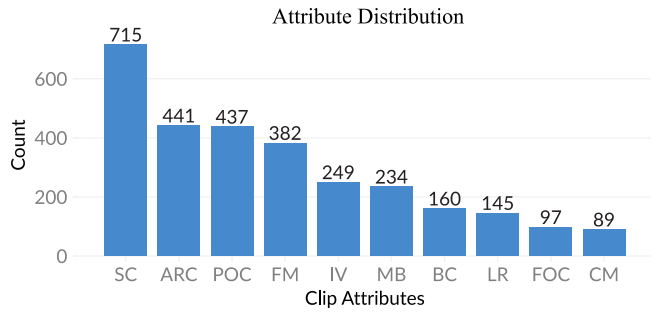


Fig. 4. Distribution of the clip attributes. The plot shows the number of single-camera (SC) clips associated with each of the attributes introduced to characterize the visual variability of the target, as in Wu et al. (2015), Fan et al. (2019), Mueller et al. (2016), Dunnhofer et al. (2023a). The application domain represented by SkiTB's videos presents a large number of scale changes (SC), followed by a substantial number of partial occlusions (POC), changes in the aspect ratio (ARC), and fast motions (FM).

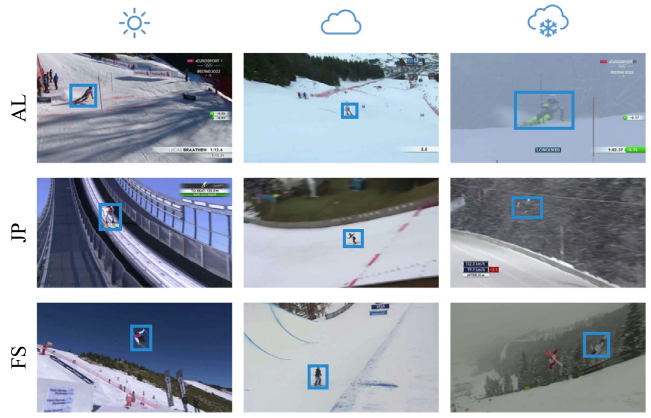


Fig. 5. Winter weather conditions. Skiing takes place in winter environments, subjecting athletes to extreme weather conditions that introduce unique image characteristics when captured on camera. For instance, "Sunny" conditions (shown in the first column of images) can create shadows, resulting in significant variations in target illumination. "Cloudy" weather (second column of images) leads to "flat light" conditions, reducing image contrast, while "Harsh" conditions such as snowfall or rain (third column) further diminishes visibility. The SkiTB includes weather condition labels for each MC video.

a fast motion (FM) situation could be present. In such a case, the performance change caused by FM would be hidden by the impact of FOC. In this view, the SC tracking setting shares similarities with the Multi-Start Evaluation protocol defined by Kristan et al. (2020), which runs a tracker at multiple points of initialization along a video to obtain more robust tracking scores. In SkiTB, the labels SC, ARC, FM, and LR, have been assigned by an automatic procedure as described by Wu et al. (2015), Fan et al. (2019). The presence of situations identified by the other attributes has been visually assessed and annotated by our research team.

The weather labels have been associated with each MC video because the weather condition generally remains the same across all the location in which the skiing competition takes place. The labeling of the conditions was performed by our team by analyzing the condition visible in the video. Such a label was also checked to match the one reported on the official result list available on the FIS database (International Ski and Snowboard Federation, 2023). The labeling generated the following weather labels: "Clouds", "Fog", "LowClouds", "Mostly-Cloudy", "Overcast", "PartlyCloud", "Raining", "Snowing", "Sunny", "Clear". In order to have a larger number of samples for the experiments, such labels have been clustered into three categories: "Sunny", "Cloudy", and "Harsh" weather. Fig. 5 gives examples of the image characteristics such weather condition cause. In total, SkiTB provides 191 videos associated with "Sunny", 66 with "Cloudy", and 43 associated with "Harsh". After the date-based training-test split, the test set used to compute the results in Fig. 10 has 80 "Sunny", 26 "Cloudy", and 14 "Harsh" videos.

4.4. Training-test splits

To enable training and evaluation of machine learning-based trackers, the MC videos are divided into training and test sets, following three different split conditions each with a 60–40 ratio. The first split follows a conventional deployment approach, where models are trained on past data and tested on newer data. This split is based on the dates associated with the videos. The second split focuses on evaluating the models' generalization ability to unseen athletes. It involves creating

Table 4
Statistics of SkiTB's training, validation, and test splits. The following table reports some statistics of the three splits that have been created to evaluate the capability of learning-based trackers in generalizing to different application conditions. For generalizing to new performances, the date associated to the videos has been used as splitting condition; for the generalization to unseen athletes, the athlete IDs; to generalize to unseen courses, the course's location information.

Generalization condition	New performances			Unseen athletes			Unseen courses		
Split	Train	Val	Test	Train	Val	Test	Train	Val	Test
# MC videos	162	18	120	155	21	124	164	18	118
# SC videos	1078	137	804	1024	133	868	1115	118	781
# frames	188 500	24 293	140 185	181 140	20 912	150 926	193 029	20 064	139 885
avg MC video seconds	39	45	39	39	33	41	39	37	40
avg SC video seconds	5.8	5.9	5.8	5.9	5.2	5.8	5.8	5.7	6.0
# sub-disciplines	11	11	11	11	11	11	11	11	11
# athletes	127	17	90	109	13	78	128	15	95
# athlete genders (M, W)	(77, 50)	(11, 6)	(53, 37)	(67, 42)	(6, 7)	(47, 31)	(76, 52)	(10, 5)	(56, 39)
# athlete nationalities	21	11	22	22	9	21	23	9	20
# locations	116	14	61	115	21	92	88	11	62
# location countries	21	8	20	23	10	23	22	9	20

separate training and test sets based on disjoint athlete IDs. The third split assesses the models' generalization to new skiing courses. In this case, dedicated disjoint partitions are formed using the location data associated with each video. Fig. 6 gives visual explanation of the composition of three different train-test partitions. The splits have been generated to maintain a balanced distribution across the skiing disciplines and sub-disciplines while aiming to keep condition-specific disjoint partitions and respect as close as possible the 60–40 ratio. The validation videos have been selected by applying the same separation strategy as previously described (but with a 90–10 ratio) to the set of training videos generated in the training-test separation phase Table 4 shows some statistics of the videos present in the three different splits.

4.5. Data quality

The video selection and annotation process was meticulously carried out by our research team, consisting of an MSc student, a post-doc researcher, and two professors. All annotators had research experience in visual object tracking and in watching skiing competitions on TV. To ensure additional accuracy, we sought application-specific guidance from two professional alpine skiing coaches and a FIS-licensed ski jumping judge. We utilized the CVAT tool (Sekachev et al., 2020) for drawing and validating the bounding-boxes. The metadata associated with the videos, including performance parameters and weather conditions, was obtained from the publicly available FIS database (International Ski and Snowboard Federation, 2023).

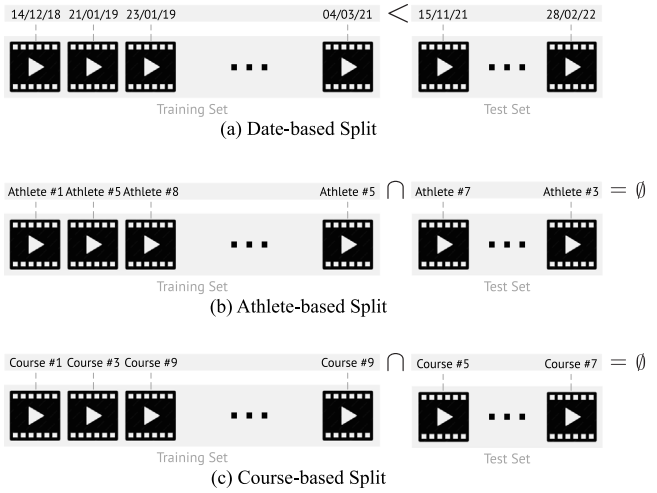


Fig. 6. Composition of the training and test video splits. For each ski discipline AL, JP, FS in SkiTB, we split the 100 MC videos into two disjoint sets, one for training machine learning models and one for testing them. (a) In the date-based split, training videos have an associated competition date that comes earlier than the date associated to test videos. (b) The athlete-based split provides training and test videos whose associated athlete IDs (athlete ID = name + surname) do not overlap. (c) The course-based split provides training and test videos where the names of the associated course locations do not overlap. After the initial training-test separation, the validation sets have been separated from the training videos by exploiting a similar approach.

5. Trackers

In this section, we give the details of the visual tracking algorithms evaluated in this study.

5.1. Generic object trackers

In our evaluation, we considered a range of state-of-the-art methods designed for tracking arbitrary objects, including long-term trackers specifically designed for addressing abrupt target changes and occlusions (Lukežič et al., 2020), as in our application of interest. The trackers falling in this category include SPLT (Yan et al., 2019), GlobalTrack (Huang et al., 2020), LTMU (Dai et al., 2020), KeepTrack (Mayer et al., 2021), STARK (Yan et al., 2021), and CoCoLoT (Dunnhofer and Micheloni, 2022; Dunnhofer et al., 2022). In addition, we included the short-term trackers (Lukežič et al., 2020) MOSSE (Bolme et al., 2010), KCF (Henriques et al., 2015), and SiamRPN++ (Li et al., 2019) for their general popularity, and MixFormer (Cui et al., 2022), OSTrack (Ye et al., 2022), FEAR (Borsuk et al., 2022), UNICORN (Yan et al., 2022), SeqTrack (Chen et al., 2023), ARTrack (Wei et al., 2023), ROMTrack (Cai et al., 2023), and ZoomTrack (Kou et al., 2023) for their very recent demonstration of high accuracy. Although these methods were not explicitly designed for long-term tracking tasks, some of them have shown promising performance in similar conditions (Fan et al., 2019) and could be even suitable for skier localization in SC tracking tasks. Table 5 reports more details on the computational architecture of the algorithms selected. All the trackers have been implemented by exploiting the code originally provided by the authors along with pre-trained weights. The original hyper-parameter values leading to the best and most likely generalizable instances of all the trackers have been set. Those trackers that do not output a confidence score, were modified to return an always-confident score of 1.0.

5.2. Skier-specific trackers

We also assessed the performance of baseline trackers specifically designed for skier tracking. This has been done to find promising research directions for the development of more accurate and robust tracking algorithms in the domain of interest.

YOLO-SORT. The YOLO-SORT tracker implements a tracking-by-detection approach inspired by MOT (Bewley et al., 2016; Dendorfer et al., 2021). At each frame of a video, this baseline first detects

Table 5
Characteristics of the generic object trackers considered in our evaluation. This table provides details about: the Image Representation employed by the trackers (Pixel - the tracker uses raw pixel intensity values; HOG - the tracker uses Histogram of Oriented Gradients; CNN architecture - the Convolutional Neural Network backbone used); the Matching operation performed to find the target in sequence frames (CF - the tracker uses correlation filters; CC - the tracker uses the cross correlation; T-by-D - the tracker uses a tracking-by-detection approach; Had - the tracker uses hadamard correlation; Tra - the tracker uses a transformer-based correlation). The ✓ symbol in the Model Update column expresses whether the tracker updates the target model during the tracking procedure. The last four columns report the category of tracking approach according to Lukežič et al. (2020) (ST_0 column - short-term trackers without any re-detection mechanism; ST_1 column - short-term trackers without any re-detection mechanism but that estimate tracking confidence; LT_0 column - pseudo long-term trackers that do not detect failure and do not perform explicit re-detection; LT_1 column - long-term trackers that detect tracking failure and perform re-detection). The evaluated long-term trackers are detailed in the first block of rows, the short-term ones in the second block.

Tracker	Venue	Image representation	Matching operation	Model	Class given by Lukežič et al. (2020)			
				Update	ST_0	ST_1	LT_0	LT_1
GlobalTrack (Huang et al., 2020)	AAAI 2020	ResNet-50	Had				✓	
LTMU (Dai et al., 2020)	CVPR 2020	ResNet-50	CF, CF, T-by-D	✓				✓
STARK (Yan et al., 2021)	ICCV 2021	ResNet-50	Tra	✓			✓	
KeepTrack (Mayer et al., 2021)	ICCV 2021	ResNet-50	CF	✓			✓	
CoCoLoT (Dunnhofer and Micheloni, 2022)	ICPR 2022	ResNet-50	CF, Tra	✓			✓	
MOSSE (Bolme et al., 2010)	CVPR 2010	Pixel	CF	✓	✓			
KCF (Henriques et al., 2015)	TPAMI 2015	HOG	CF	✓	✓			
SiamRPN++ (Li et al., 2019)	CVPR 2019	ResNet-50	CC		✓			
MixFormer (Cui et al., 2022)	CVPR 2022	Custom Transformer	Tra	✓		✓		
OSTrack (Ye et al., 2022)	ECCV 2022	ViT	Tra		✓			
FEAR (Borsuk et al., 2022)	ECCV 2022	ResNet-50	CC	✓	✓			
UNICORN (Yan et al., 2022)	ECCV 2022	ConvNeXt	Tra	✓		✓		
SeqTrack (Chen et al., 2023)	CVPR 2023	ViT	Tra	✓		✓		
ARTrack (Wei et al., 2023)	CVPR 2023	ViT	Tra	✓	✓			
ROMTrack (Cai et al., 2023)	ICCV 2023	ViT	Tra	✓	✓			
ZoomTrack (Kou et al., 2023)	NeurIPS 2023	ViT	Tra	✓	✓			

skiers with an YOLOX instance (Ge et al., 2021) and then exploits the Simple Online and Realtime Tracking method (SORT) (Bewley et al., 2016) to associate the new detections with previously memorized tracklets (i.e. sequences of bounding-boxes referring to the same target). The YOLOX instance was trained on all the frames and the associated bounding-box annotations of SkiTB's combination of training and validation sets defined by the date-based split, by mostly default hyper-parameters. The only changes made are relative to the batch size, set to 16, and the number of training epochs, set to 25. 10% of the training videos were considered to build the set of validation images. The model instance achieving the highest Average Precision (AP) on such a subset was retained for inference during tracking. For each video, the SORT module is initialized in the first frame with the given skier's bounding-box. At every other frame, the module is given in input all the detections given by YOLOX and returns a new set of tracks. As output, we retain the bounding-box associated with the track initialized in the first frame.

STARK$_{\texttt{FT}}$. The STARK$_{\texttt{FT}}$ baseline implements a fine-tuned version of the generic object tracker STARK (STARK-ST50) (Yan et al., 2021). To implement this tracker we exploited the publicly available code and adapted the model's tracking ability by fine-tuning on SkiTB's combination of training and validation sets, according to STARK's original training strategy. Mostly default hyper-parameters have been kept, except for the number of epochs in stage-one training, which has been set to 200. During inference, this baseline acts in the same ways as the original STARK.

STARK$_{\texttt{SKI}}$. Furthermore, we introduce a new and better-performing tracker, called STARK$_{\texttt{SKI}}$. The pseudo-code of the procedure implemented by this skier-optimized baseline for an MC video is given in Algorithm 1. In simple words, the procedure is composed of two skier-specific instances of STARK$_{\texttt{FT}}$. The first one, which we refer to as STARK$_{\texttt{FT-SC}}$, is a modified version of STARK$_{\texttt{FT}}$ that, at every frame, computes the target bounding-box by exploiting a higher-resolution search area located around the previous target location. This is achieved by reducing the search area factor from the original value of 5.0 to 3.0 (we determined the value 3.0 by experiments) and fine-tuning as done for STARK$_{\texttt{FT}}$. In this way, we reduce the amount of background information present in the search area, thus increasing

Algorithm 1 Pseudo-code of the procedure implemented by the proposed STARK$_{\texttt{SKI}}$ while running on a video.

1: *// Consider video $\mathcal{V}$ and ground-truth box b_0*
2: *// Trackers initialization*
3: Initialize STARK$_{\texttt{FT-SC}}$with F_0 and b_0
4: Initialize STARK$_{\texttt{FT}}$with F_0 and b_0
5: $t \leftarrow 1$
6: **repeat**
7: $b_t, c_t \leftarrow$ Run STARK$_{\texttt{FT-SC}}$ on F_t
8: **if** $c_t \leq \delta$ **then**
9: $b_t, c_t \leftarrow$ Run STARK$_{\texttt{FT}}$ on F_t
10: **if** $c_t > \delta$ **then**
11: *// STARK$_{\texttt{FT-SC}}$ re-initialization*
12: Re-initialize STARK$_{\texttt{FT-SC}}$ with F_t and b_t
13: **end if**
14: **else**
15: *// Compute bounding-box for STARK$_{\texttt{FT}}$ relocalisation*
16: $S \leftarrow \frac{H}{5.0}$ // 5.0 is STARK$_{\texttt{FT}}$ search area's factor
17: $x_t^* \leftarrow clip(x_t, \frac{H}{2}, W - \frac{H}{2})$
18: $y_t^* \leftarrow \frac{H}{2} - \frac{S}{2}$
19: $b_t^{(R)} \leftarrow [x_t^*, y_t^*, S, S]$
20: Use $b_t^{(R)}$ to reset STARK$_{\texttt{FT}}$'s box used to compute the search area location
21: **end if**
22: Return b_t, c_t as output for F_t
23: $t \leftarrow t + 1$
24: **until** $t = T$

the resolution of the target skier's appearance and making the tracker predict more accurate bounding-boxes during SC tracking. Given the more limited search area, STARK$_{\texttt{FT-SC}}$ performs better just in such conditions where the target and camera motion are stable and consistent across consecutive frames. In the other cases, i.e. in those frames where STARK$_{\texttt{FT-SC}}$ is not confident in tracking the target (i.e., when the STARK$_{\texttt{FT-SC}}$'s confidence score $c_t \leq \delta$, lines 8–13 of Algorithm 1), we exploit a STARK$_{\texttt{FT}}$ instance configured as described in the previous paragraph. This instance keeps the original search factor with a value of 5.0 and thus is able to look for the target in a larger frame area. The execution of this STARK$_{\texttt{FT}}$'s instance is generally triggered after a camera shot-cut and during the complete occlusion of the target. We

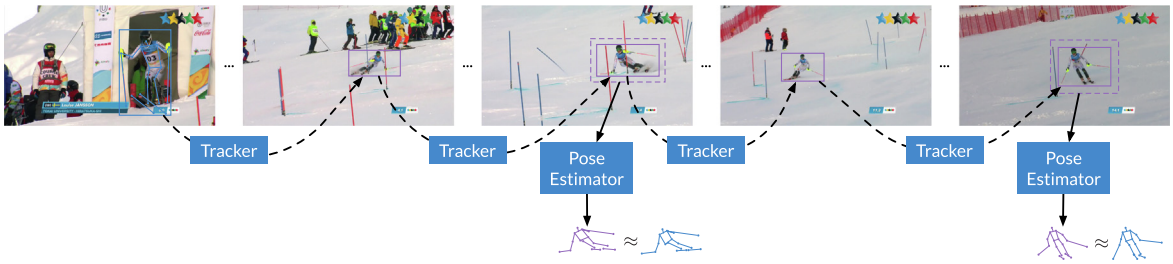


Fig. 7. Evaluation protocol to quantify the tracking impact. This figure visualizes the procedure we implemented to evaluate a visual tracker based on the performance of an high-level skiing performance understanding module such as a 2D pose estimator (e.g., AlphaPose). On the Ski2DPose and YouTube Skijump dataset, we consider each sequence of frames with sparsely annotated 2D body and equipment poses. A tracker is initialized in the frame with the first available pose, from which a bounding-box is computed (□). The tracker is then run on each subsequent frame to provide the bounding-box referring to the same skier (□). At each frame with the ground-truth pose, a larger patch (dashed □) is extracted from the tracker's box and it is given in input to the pose estimation algorithm. The latter returns as output the 2D pose of the skier, which is compared with the ground-truth. The average value of such a distance computed across pose-annotated frame is used to form a ranking of different trackers that expresses their impact on the quality of the poses.

empirically found it beneficial to set the search area size of this instance to match the frame's height, by modifying the bounding-box values that are used to compute the search area at the next frame (lines 16–20 of Algorithm 1). The position of such a box is set to be the latest confident box position predicted by $\text{STARK}_{\text{FT-SC}}$, clipped to make the search area not fall outside of the frame. Whenever STARK_{FT} finds confidently the target again (i.e., with $c_t > \delta$), its predicted bounding-box and the respective frame are used to re-initialize $\text{STARK}_{\text{FT-SC}}$. We found the re-initialization to work better than just relocating $\text{STARK}_{\text{FT-SC}}$ on the STARK_{FT}'s predicted bounding-box. We found empirically $\delta = 0.5$ to act well for the thresholds of both $\text{STARK}_{\text{FT-SC}}$'s and STARK_{FT}'s c_t scores.

6. Evaluation

In this section, we explain and motivate in detail the evaluation procedures implemented to quantify the tracking accuracy and efficiency, and the impact of the trackers on the accuracy of high-level skiing performance understanding tasks. If not specified otherwise, in the experiments all the trackers were executed on the date-based test-set of SkiTB, with skier-specific trackers trained on the corresponding training set.

6.1. Tracking performance

Evaluation protocol. To run a tracker for evaluation of its tracking accuracy, we employed the one-pass evaluation (OPE) protocol introduced by Wu et al. (2015) which implements the most realistic way to run a tracker in practice. The protocol consists of two main stages: (i) initializing a tracker with a bounding-box of the target in the first frame of the video; (ii) letting the tracker run on every subsequent frame until the end and recording bounding-box predictions to be considered for the evaluation. To obtain performance scores for each sequence, predictions and ground-truth bounding-boxes are compared according to some distance measure. The overall scores are obtained by averaging the scores achieved for every sequence. As in the default OPE, we use the ground-truth bounding-box for initialization to evaluate the trackers in the best possible conditions, i.e. when accurate information about the target is given. However, many deployment conditions do not allow human labeling but instead require a completely automatic athlete localization system (e.g. real-time skiing performance analysis during broadcasting). To evaluate trackers in similar conditions, we use an object detector to predict the initial skier bounding box. Thus, we consider a version of the OPE protocol where each tracker is initialized in the first frame in which the YOLOX detector's (Ge et al., 2021; Jocher et al., 2020), fine-tuned for skier localization, provides a bounding-box prediction with confidence score ≥ 0.5. The fine-tuning of this detector was performed in the same way as for YOLO-SORT baseline described before.

Performance measures. To quantify the distance between the predicted and temporally-aligned ground-truth bounding-boxes, we used different measures. As general tracking accuracy indicators, we employed the metrics defined by Lukežič et al. (2020) for long-term tracking problems: Precision, Recall, and F-Score. Due to the generally long video observation and presence of multiple occlusions, our problem of interest is related to such a research framework. The Precision (Pr ↑) measures the average amount of correctly tracked ground-truth bounding-boxes where the tracker is confident, with different thresholds used to determine the conditions of correct and confident prediction. In the case of our domain, the Pr ↑ score determines the average coverage of the skier's position on the portion of skiing performance observation on which the tracker is confident. For example, a Pr ↑ score of 0.8 tells that an algorithm correctly localizes the athlete for 80% of the bounding-box predictions that are given with high confidence. The Recall (Re ↑) instead measures the average amount of correctly tracked ground-truth bounding-boxes, regardless of the tracker's confidence. In our context, such a score determines the average coverage of the position of the skier throughout the whole skiing performance. For instance, a Re ↑ score of 0.8 gives that the algorithm correctly localizes the athlete for 80% of the skiing performance appearing in the video. The F-Score (F-Score ↑) provides a single aggregating score that incorporates both the previous measures. The best value across the different confidence thresholds is retained.

In addition to those metrics, we exploited the Generalized Success Robustness (GSR ↑) (Dunnhofer et al., 2023a) which reports the fraction of continuous successful tracking before the tracker is lost, measured as the temporal index of the first wrong prediction normalized by the number of frames in the video. In the context of this application domain, such a metric reports the percentage of continuous coverage of the skier's performance before the target is lost by the tracking algorithm. The original metric (Dunnhofer et al., 2023a) is strict because it considers just the first wrong prediction to determine the tracker's failure time step. Other work (Kristan et al., 2020) suggested a softer version of such a measure. If the algorithm gets back to the target within a range of 10 consecutive frames, the tracking is resumed. Inspired by such a work, we evaluate the GSR ↑ with several different temporal ranges to detect a failure, specifically 1 frame (∼0.03s), 7 frames (∼0.25s), 15 frames (∼0.5s), 22 frames (∼0.75s), 30 frames (∼1s), 60 frames (∼2s), and 90 frames (∼3s).

Finally, we assessed the computational efficiency of the trackers. This has been done by quantifying the time difference (in seconds) between the time stamp associated with each frame and the time instant on which the localization for the respective frame is given by an algorithm. Considering that sports performance analysis requires the processing of all the frames for a smooth and continuous understanding, a tracker that is slow will accumulate time while processing all the frames and delay its predictions. Thus, it becomes interesting to know how much time should be waited in order to obtain the localization,

and how such delay grows during the online processing of the video. We give such a measurement in seconds with Delay ↓. In addition to such an evaluation methodology, we also employed the number of frames-per-second (FPS) and the number of Floating Point Operations (FLOPs) to assess the efficiency and the complexity of the tracking methods.

6.2. Tracking impact

The output of tracking is of paramount importance for many high-level modules that produce fine-grained skiing performance analyses (Štepec and Skočaj, 2022; Wang et al., 2019; Ludwig et al., 2023, 2022; Dunnhofer et al., 2023b). Thus we evaluated the trackers based on the impact they have on the accuracy of such solutions. We think that the development of effective tracking methodologies should be driven not only by tracking performance results but also by the contribution the algorithms bring to improving the accuracy of the overall systems.

As examples of high-level skiing performance understanding tasks to evaluate trackers' impact, we focused on the problem of 2D pose estimation of skier body and equipment (Bachmann et al., 2019; Ludwig et al., 2023). Solving this task serves to obtain information regarding the position and orientation of specific human joints during exercises, and such an output is additionally exploited by even more high-level performance understanding modules such as 3D pose estimation (Bachmann et al., 2019; Wandt et al., 2021). To estimate the image-level coordinates of a set of key-points that localize different parts of a skier's body (e.g. head, shoulders, hips, feet, etc.) and of particular points of interest of the skier's equipment (e.g. ski tips or tails), the available solutions (Bachmann et al., 2019; Ludwig et al., 2023) first run an object detector (Ren et al., 2015; Redmon et al., 2016) to compute bounding-boxes for the athlete present in the input RGB image, and then crop image patches from such boxes that are successively given as input to a state-of-the-art deep neural network architecture – e.g. AlphaPose (Fang et al., 2022) – that predicts the key-point coordinates. Such a pose estimation network is trained by fine-tuning on ground-truth poses by exploiting input image patches extracted with bounding-boxes defined by the coordinates of the annotated key-points.

The aforementioned studies (Bachmann et al., 2019; Ludwig et al., 2023) propose the datasets of videos Ski2DPose dataset (Bachmann et al., 2019) (for the AL discipline) and YouTube Skijump dataset (Ludwig et al., 2023) (for the JP discipline), with dedicated training and test sets, that have sparse frames labeled with the poses of body and ski equipment. The authors evaluate their proposed pipelines on such benchmarks but treat each frame as an independent image, thus during testing an object detector is run on every image before the pose estimation network. Considering the presence of videos, we use such datasets as a base for the evaluation of trackers as athlete localizers before the pose estimation step. Hence, we determine the tracker's impact by evaluating the accuracy of the pose estimation model, where the input of the latter is influenced by the output of the former. After having fine-tuned an AlphaPose instance (Fang et al., 2022) on the original training images as used by Bachmann et al. (2019), Ludwig et al. (2023), we evaluate its accuracy on the relative test frames by inputting it with a patch extracted from a tracker's box prediction. The evaluation of the pose estimator is done through: the Percentage of Correct Key-points (PCK ↑) which measures the number of predicted key-points, normalized by the number of all key-points (Bachmann et al., 2019), having a pixel distance lower than the 50% of the ground-truth-based head-neck distance; and the Mean Per Joint Position Error (MPJPE ↓) which measures the normalized pixel distance between predicted and corresponding ground-truth key-points (Bachmann et al., 2019). The tracker's bounding-boxes are obtained by implementing the OPE protocol on the sequence of frames in between the first and the last pose annotation occurrences that refer to the same athlete. Indeed, we obtain boxes's top-left and bottom-right vertices by considering the lowest and greatest values in the key-points coordinates. The first bounding-box is considered for tracker initialization, while the others are for evaluation according to the metrics defined in the previous sub-section. A visual representation of such an evaluation protocol is presented in Fig. 7. We respect the original training-test separations (Bachmann et al., 2019; Ludwig et al., 2023). For testing on the Ski2DPose dataset, we used 11 video clips related to the 150 pose annotated images, while for the YouTube Skijump dataset we used 19 videos built upon the 118 annotated validation images.

For the implementation and fine-tuning of the AlphaPose instance (Fang et al., 2022), we employed the Alphapose v0.6 framework based on the ResNet50 model. Specifically, we conducted two separate fine-tuning for Ski2DPose and YouTube Skijump. Both training sessions ran for 250 epochs, employing a batch size of 32 and a learning rate of 0.001 decreased by a 0.1 factor every 70 epochs. During both training and testing, in the computation of the input image crop, a padding of 20% was added to the dimensions of the available bounding-box.

6.3. Implementation details

All the code used for this study was implemented in Python and run on a machine with an Intel Xeon E5-2690 v4 @ 2.60 GHz CPU, 320 GB of RAM, and 8 NVIDIA TITAN V GPUs.

7. Results

In this section, we report on the outcomes of the conducted study.

7.1. General tracking performance

Multi-camera tracking. Table 6 presents the tracking performance expressed by the F-Score ↑, Pr ↑, Re ↑ metrics achieved in the MC setting by all selected trackers on the date-based test set of SkiTB. Among generic object trackers, STARK results the best. In terms of absolute values, its performance is comparable to the one achieved on traditional long-term benchmark datasets (Kristan et al., 2021; Fan et al., 2019). We also observe that long-term trackers (STARK, CoCoLoT, LTMU) surpass the more recent methodologies based on transformers (OSTrack, SeqTrack, MixFormer) that are not explicitly designed for such a setting. Unified methodologies for object localization tasks (e.g., UNICORN) exhibit promising performance in MC conditions. Leveraging object detection training could be beneficial for re-detecting the target. However, UNICORN does not surpass tracking algorithms specifically designed for long-term tracking problems. In general, these results indicate that generic object trackers struggle to generalize to the application settings represented by SkiTB. These findings highlight that generic object trackers are not yet suitable for deployment in computer vision systems aimed at understanding skiing performance comprehensively, covering the entire duration of the performance from the beginning to the end. The observed low tracking accuracy has the potential to negatively impact over 40% of the results in the analysis.

On the other hand, skier-specific trackers (YOLO-SORT, STARK_{FT}, $\text{STARK}_{\text{SKI}}$) perform significantly better, with STARK_{FT} improving STARK's F-Score ↑ by 40%. In the same score, $\text{STARK}_{\text{SKI}}$ achieves an additional 2% increase over STARK_{FT}. We observe that $\text{STARK}_{\text{SKI}}$ achieves a higher score in Pr ↑ rather than in Re ↑. This result indicates that the algorithm is accurate at 84.3% in localizing the target skier when the estimated target presence is high, while it is accurate at 82.9% when the target is actually present. Even though the difference between the two numbers is minimal, these findings suggest that the bounding-box prediction of the algorithm is more accurate than the target presence prediction. Overall, skier-specific trackers exhibit much more promising performance in consistently tracking the skier's appearance throughout the entire skiing performance. But the results imply that the best-performing method, $\text{STARK}_{\text{SKI}}$, is not flawless, and approximately 17% of the skiing performance understanding analysis might be affected by potentially incorrect target skier localizations.

Table 6
Overall and per-discipline results in the multi-camera (MC) tracking setting. The F-Score ↑, Pr ↑, and Re ↑ scores are presented for each studied algorithm. Best, second-best, and third-best overall scores are highlighted in **gold**, **silver**, and **bronze**, respectively. **Bold** highlights the best results among generic object trackers. The latters struggle to keep track of the skiers in this setting, while skier-specific trackers demonstrate promising capabilities. Application-wise, we observe that ski jumping (JP) is the discipline in which trackers perform better, followed by alpine skiing (AL). Freestyle skiing (FS) offers the most challenging situations.

Discipline	MOSSE	KCF	SiamRPN++	FEAR	GlobalTrack	ROMTrack	ARTrack	MixFormer	KeepTrack	OSTrack	SeqTrack	ZoomTrack	LTMU	UNICORN	CoCoLoT	STARK	YOLO-SORT	$STARK_{FT}$	$STARK_{SKI}$
All	0.093	0.061	0.248	0.338	0.493	0.494	0.510	0.526	0.527	0.528	0.534	0.546	0.554	0.559	0.562	**0.584**	**0.740**	**0.818**	**0.835**
	0.092	0.061	0.270	0.419	0.493	0.487	0.503	0.518	0.555	0.520	0.538	0.538	0.565	0.559	0.572	**0.595**	**0.730**	**0.832**	**0.843**
	0.094	0.062	0.235	0.301	0.495	0.503	0.519	0.535	0.508	0.537	0.533	0.556	0.545	0.560	0.555	**0.576**	**0.751**	**0.806**	**0.829**
AL	0.031	0.024	0.144	0.270	0.485	0.430	0.452	0.463	0.518	0.462	0.479	0.487	0.524	0.532	0.532	0.552	0.798	0.853	0.868
	0.031	0.024	0.143	0.260	0.487	0.426	0.447	0.458	0.561	0.457	0.485	0.482	0.541	0.537	0.546	0.565	0.790	0.874	0.885
	0.032	0.024	0.145	0.229	0.483	0.435	0.483	0.456	0.484	0.467	0.475	0.492	0.509	0.526	0.521	0.540	0.807	0.834	0.852
JP	0.155	0.098	0.281	0.373	0.504	0.535	0.546	0.574	0.536	0.577	0.590	0.586	0.576	0.598	0.584	0.603	0.818	0.880	0.896
	0.153	0.097	0.310	0.451	0.507	0.529	0.540	0.567	0.576	0.571	0.598	0.580	0.591	0.602	0.606	0.630	0.807	0.892	0.898
	0.157	0.099	0.262	0.338	0.502	0.542	0.553	0.581	0.510	0.584	0.584	0.594	0.565	0.596	0.569	0.582	0.830	0.871	0.896
FS	0.092	0.065	0.319	0.372	0.491	0.517	0.533	0.541	0.528	0.545	0.533	0.566	0.562	0.547	0.570	0.596	0.603	0.721	0.742
	0.090	0.067	0.358	0.446	0.483	0.505	0.521	0.529	0.528	0.532	0.530	0.530	0.552	0.538	0.564	0.590	0.592	0.730	0.746
	0.094	0.080	0.298	0.336	0.500	0.531	0.548	0.556	0.530	0.560	0.539	0.581	0.562	0.557	0.577	0.604	0.616	0.713	0.738

Table 7
Overall and per-discipline results in the single-camera (SC) setting. The F-Score ↑, Pr ↑, and Re ↑ scores are presented for each studied algorithm. Best, second-best, and third-best overall scores are highlighted in **gold**, **silver**, and **bronze**, respectively. **Bold** highlights the best results among generic object trackers. This setting is easier to tackle by all the algorithms in general. Generic trackers perform much better in this scenario than in the multi-camera (MC) videos, but skier-specific trackers still perform better. The different skiing discipline pose challenges to the trackers in the same way as in the MC setting.

Discipline	KCF	MOSSE	FEAR	SiamRPN++	GlobalTrack	UNICORN	LTMU	SeqTrack	ARTrack	KeepTrack	OSTrack	ZoomTrack	CoCoLoT	MixFormer	ROMTrack	STARK	YOLO-SORT	$STARK_{FT}$	$STARK_{SKI}$
All	0.294	0.367	0.564	0.583	0.592	0.613	0.642	0.645	0.651	0.654	0.663	0.680	0.681	0.686	0.699	**0.703**	**0.751**	**0.836**	**0.841**
	0.291	0.363	0.565	0.583	0.591	0.607	0.637	0.639	0.642	0.652	0.654	0.670	0.676	0.676	0.689	**0.698**	**0.743**	**0.827**	**0.833**
	0.299	0.376	0.572	0.592	0.601	0.626	0.651	0.681	0.667	0.664	0.704	0.698	0.694	0.658	0.717	**0.717**	**0.763**	**0.854**	**0.858**
AL	0.220	0.267	0.518	0.536	0.585	0.572	0.623	0.578	0.605	0.640	0.594	0.627	0.652	0.637	0.651	0.671	0.819	0.875	0.882
	0.218	0.265	0.524	0.542	0.595	0.570	0.622	0.576	0.600	0.641	0.590	0.623	0.650	0.634	0.646	0.672	0.814	0.876	0.886
	0.222	0.269	0.516	0.534	0.578	0.575	0.625	0.580	0.609	0.640	0.599	0.633	0.655	0.643	0.656	0.672	0.825	0.876	0.881
JP	0.389	0.487	0.641	0.663	0.677	0.705	0.702	0.747	0.736	0.705	0.763	0.756	0.738	0.765	0.782	0.761	0.855	0.899	0.907
	0.388	0.485	0.645	0.666	0.677	0.704	0.703	0.748	0.733	0.707	0.760	0.753	0.743	0.762	0.779	0.766	0.850	0.894	0.901
	0.390	0.489	0.640	0.660	0.678	0.708	0.701	0.747	0.740	0.703	0.767	0.761	0.734	0.770	0.786	0.758	0.863	0.907	0.914
FS	0.274	0.347	0.531	0.550	0.514	0.563	0.599	0.612	0.612	0.617	0.633	0.655	0.654	0.654	0.663	0.676	0.578	0.734	0.735
	0.268	0.338	0.525	0.540	0.502	0.547	0.586	0.595	0.593	0.607	0.612	0.634	0.636	0.633	0.642	0.656	0.566	0.711	0.713
	0.285	0.370	0.560	0.582	0.548	0.595	0.627	0.647	0.652	0.647	0.676	0.701	0.692	0.698	0.709	0.721	0.601	0.780	0.780

Single-camera tracking. Comparing the latter findings with the results obtained for the SC setting available in Table 7, we observe that camera shot-cuts introduce challenges that adversely affect the tracking performance of all the methods, skier-specific and non. The SC setting resembles the problem of short-term visual object tracking (Wu et al., 2015; Lukežič et al., 2020) where the videos are captured by the same video camera and their duration is up to few seconds. Application-wise, the conditions of SC tracking align with: the replays during broadcasting transmission where just a specific section of the skiing performance is captured by a single camera and played again later; videos acquired during training processes where a trainer captures a specific section of the ski track/course with a smartphone for later video analysis. From the table, we observe that generic object trackers show a larger improvement in tracking performance by working on SC videos rather than on MC ones. The performance of STARK improves by 20% in this setting, achieving values that represent promising tracking performances for application (Kristan et al., 2021; Fan et al., 2019). In this scenario, short-term trackers such as ROMTrack, MixFormer, OSTrack surpass long-term ones like CoCoLoT and KeepTrack. Nevertheless, the tracking accuracy of all the generic object tracking algorithms still remains lower than the skier-specific methods. $STARK_{SKI}$ keeps the top spot in the ranking, with a 0.6% improvement over $STARK_{FT}$ and 12% over YOLO-SORT. The difference in F-Score ↑ between $STARK_{SKI}$ running in MC and SC conditions is merely 0.7%. This result implies that, while camera shot-cuts do influence tracking behavior, the accuracy of the skier-specific tracker remains affected by other domain characteristics that hinder a perfectly accurate learning of the skier's appearance and motion patterns, as it has been showcase for MC videos in the previous paragraph.

Length of temporal reference. Fig. 8 displays the proportion (as GSR ↑ scores) of the skiing performance that the trackers are able to consistently cover before losing track of the target skier, starting from the beginning of the skier's performance. Overall, the skier-specific trackers show to be able to keep a longer reference to the target than generic object methods. $STARK_{SKI}$ demonstrates the most promising result. With a recovery time of 1 s or longer, fractions scores exceed 80%. In the same setting, the best generic object tracker, revUNICORN, achieves values around 70%. Shorter recovery time thresholds result in shorter target coverages. With a threshold of 1 frame (i.e., 0.03s), $STARK_{SKI}$ achieves successful continuous tracking for more than the first 40% of the athlete's performance. But this result is still better than UNICORN which results in a coverage of around 25%. These results illustrate that, even with skier-specific trackers, achieving consistently accurate per-frame skier localization is not yet realized. In other words, at the current state-of-the-art, incorrect and abrupt bounding-box predictions may temporarily impact the output of computer vision systems utilizing skier localization from these trackers. Although the trackers generally recover automatically after a few seconds, real-time continuously-performing systems might experience negative effects that should be mitigated with some filtering operations.

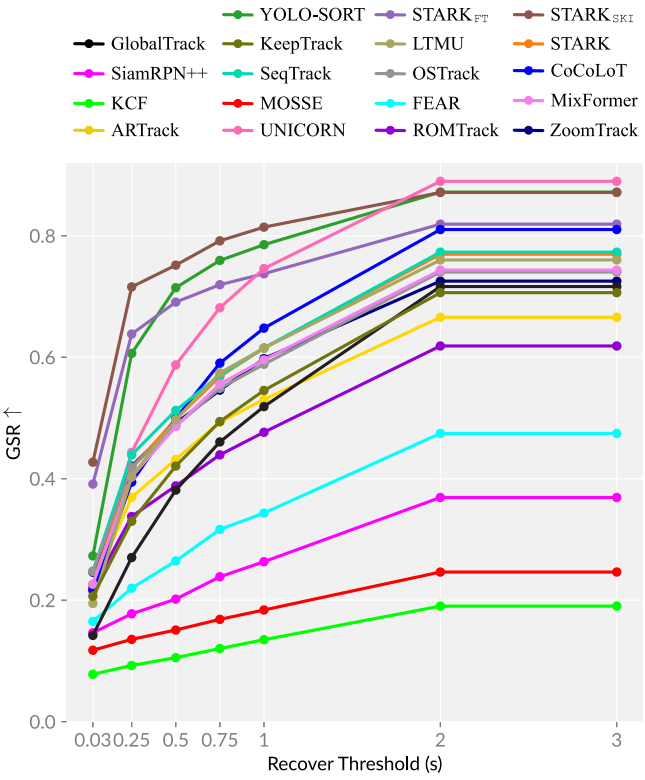


Fig. 8. Fraction of consistent skier tracking (length of temporal reference) starting from the beginning of the ski performance. These plots depict the average fraction of consecutive frames in which the target skier is accurately localized before losing track, as measured by the GSR score (Dunnhofer et al., 2023a). Various time thresholds in seconds are employed to assess the trackers' ability to recover from failures over time (Kristan et al., 2020). The plot reports the results obtained by both skier-specific trackers and by the generic object trackers.

Attribute-dependent tracking performance. By analyzing the F-Score ↑ per the visual attributes characterizing the SC clips, as reported by Fig. 9, we notice that the full occlusion (FOC), the small size (LR), and the fast motion (FM) of skiers are the conditions that determine a performance drop to both skier-specific and generic object trackers. It is

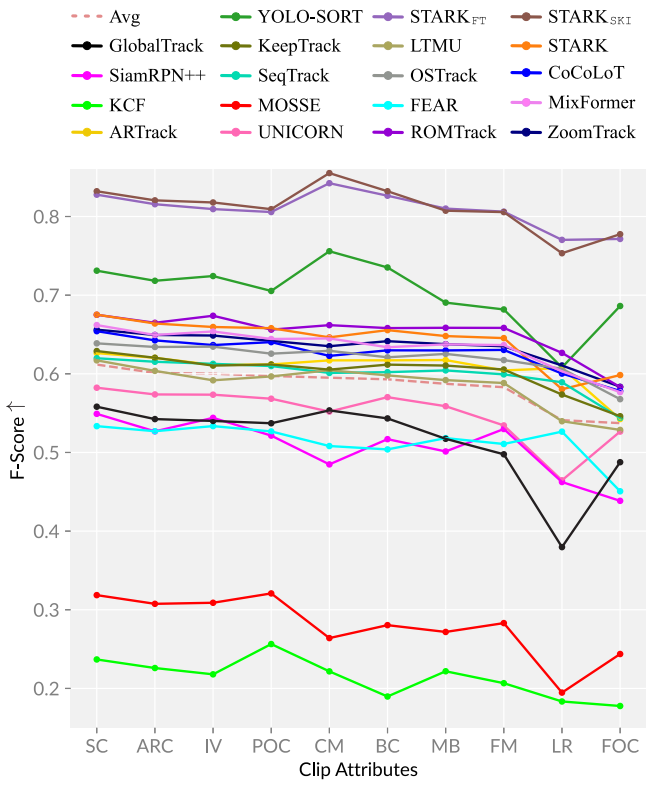


Fig. 9. Tracking performance based on visual attributes. This plot reports the F-Score ↑ for the different attributes used to characterize the single-camera (SC) clips. We observe that the low resolution (LR), the full occlusion (FOC), and the fast motion (FM) of skiers are the most difficult situations to address for all the trackers. The plot reports the results obtained by both skier-specific trackers and the generic object trackers.

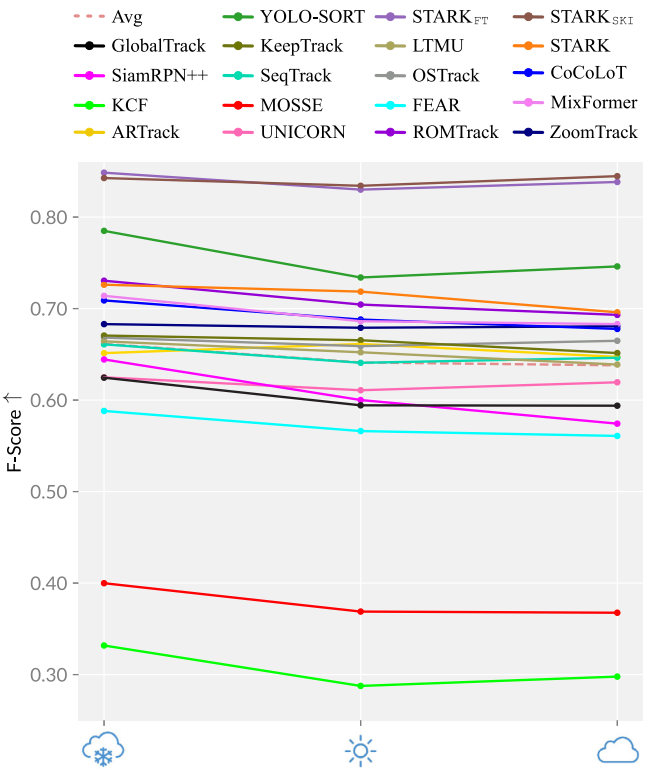


Fig. 10. Impact of the weather conditions on tracking. This plot displays the F-Score ↑ results conditioned on the weather conditions (harsh, sunny, cloudy) characterizing the SC clips. Plot reports both the skier-specific trackers and the generic object ones. In general, we observe that the tracking accuracy is not influenced much by different weather conditions.

worth noticing that, the order of the LR and FOC attributes (that is obtained by ordering in decreasing order the average F-Score ↑) changes between skier-specific and generic trackers. This result suggests that fine-tuning on skier appearance helps in resolving complete occlusions of the target skiers. However, when applying these trackers in real-world conditions inside computer vision systems, we recommend, if feasible, avoiding situations that might pose the most severe challenges. This can be achieved by carefully controlling the position and zoom level of the camera relative to the skiing course. On the other hand, situations involving scale change (SC), aspect ratio change (ARC), and illumination variation (IV) are better addressed on average. Skier-specific trackers demonstrate greater proficiency in handling camera motion (CM) and background clutter (BC) compared to generic object trackers. In deployment conditions, the presence of these scenarios can be expected to exert limited influence on the accuracy of the skier localization.

Weather-dependent tracking performance. The plots in Fig. 10 show that both the skier-specific and generic object trackers work generally with the same level of accuracy under the different weather conditions considered in SkiTB. The $\text{STARK}_{\text{SKI}}$ and STARK_{FT} trackers demonstrate a more balanced tracking accuracy between "Harsh", "Sunny", and "Cloudy" weather conditions. YOLO-SORT is more susceptible to the last two conditions. Surprisingly, the trackers tend to perform better in conditions of challenging weather conditions, while clear weather conditions impact slightly more the tracking performance. This outcome suggests that the high shadowing present in sunny weather and the flat light in cloudy conditions are slightly more difficult to cope with. Overall, these results demonstrate that, in scenarios of broadcasting videos of professional athletes as represented by SkiTB, where camera objectives are quite clear from snow, rain, or fog, the skier-specific methodologies can be used reliably even in the case of challenging image conditions caused by weather conditions.

Table 8
Detector-based initialization. The F-Score ↑ of different trackers is compared in terms of a ground-truth-based (left of →) and detection-based initialization (right of →). Overall, the skier-specific trackers show to be robust to the noise in the bounding-box used for initialization. The least performance drop, achieved with YOLO-SORT, is highlighted in **bold**.

Tracker	AL	JP	FS	All
STARK	0.552 → 0.544	0.603 → 0.590	0.596 → 0.497	0.584 → 0.544
YOLO-SORT	0.798 → 0.798	0.818 → 0.818	0.603 → 0.588	**0.740 → 0.735**
STARK_{FT}	0.853 → 0.850	0.880 → 0.879	0.721 → 0.698	0.818 → 0.809
$\text{STARK}_{\text{SKI}}$	0.868 → 0.870	0.896 → 0.897	0.742 → 0.696	0.835 → 0.821

Qualitative results. Fig. 12 gives visual illustration of the conclusions made in the previous paragraph for of the top four methods STARK, YOLO-SORT, STARK_{FT}, $\text{STARK}_{\text{SKI}}$, in the MC tracking setting. Overall, we can state that skier-specific trackers show promising performance for the application in real-world, especially in videos acquired by the same camera. Fig. 13 instead shows qualitative examples in the case of the particular image conditions depending on the weather. Fig. 14 depicts video frames of complex situations such as the low resolutions of targets (identified by the LR attribute), the complete occlusions of the skier (identified by the FOC attribute), the presence of skiers with similar appearance (i.e., distractors as present in the FS sub-discipline of ski cross), that make the best trackers fail.

7.2. In-depth analysis

In this section, we present a deeper analysis of the application domain's impact on the four most accurate methods presented in Table 6, namely STARK, YOLO-SORT, STARK_{FT}, and $\text{STARK}_{\text{SKI}}$. We also assess

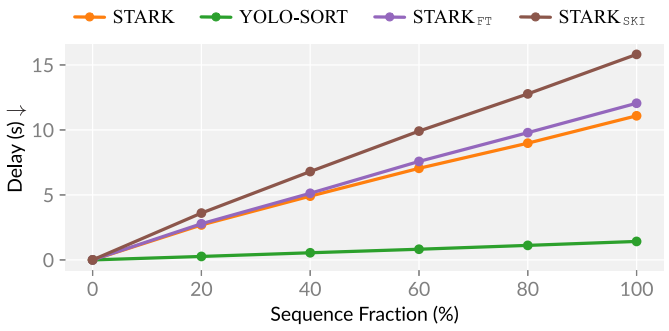


Fig. 11. Waiting time to obtain skier localizations. The plot illustrates at various fractions of an MC sequence the average time that has to be waited to get the bounding-boxes from the trackers. YOLO-SORT demonstrates the highest efficiency, with minimal delay compared to the actual happening of the skiing performance.

the design and training choices' effects of our proposed STARK$_{FT}$ and STARK$_{SKI}$ baselines on the tracking accuracy.

Ski-discipline-dependent tracking performance. From both Tables 6 and 7 it can noticed that ski jumping (JP) receives the best tracking performance among the skiing disciplines, followed by alpine skiing (AL), while freestyle skiing (FS) lead to more challenging situations. With respect to AL, the performance of STARK$_{SKI}$ is reduced by 17%. This may be attributed to the complex poses athletes perform in such a discipline. Indeed, as evidenced by the last two rows of frames in Fig. 12, FS athletes perform aerial maneuvers that significantly alter their appearance in videos. Frequently, the intricate poses assumed during these maneuvers can make the athlete's appearance resemble other objects in the scene, leading to instances of background clutter and potential confusion with other objects. Furthermore, FS videos exhibit characteristics such as the presence of other skiers (as observed in the sub-disciplines of ski cross and dual moguls, depicted in the last row of frames in Fig. 14) who perform concurrently with the target skier. In such scenarios, the target athlete may be occluded by other skiers, leading to potential confusion with athletes wearing similar suit colors or graphics. Addressing this challenge may require the development of tracking algorithms capable of discerning very subtle appearance differences to accurately distinguish one skier from another.

Initialization with a detector. Table 8 presents the impact, as reflected in the change in F-Score ↑, of using the YOLOX detector (Ge et al., 2021; Jocher et al., 2020) to initialize the tracker. It is observed that the generic STARK is more sensitive to initialization noise as it loses around 7% in F-Score ↑. Given that tracking the appearance of both the skier's body and skis might not be a common instance in the training videos of generic object trackers, introducing noise into the initialization box could further include background pixels that impact the extraction of representative target features. This, in turn, may hinder the precision of localization while predicting bounding-box coordinates in successive video frames. Conversely, skier-specific methods exhibit greater robustness to such initialization, with STARK$_{SKI}$ maintaining good scores across all the disciplines and dropping the F-Score ↑ by only 2% overall. Results from the YOLO-SORT skier-specific methodology experience minimal impact from noisy initialization information. This is attributed to the methodology's application of a previous-frame-independent target localization, which does not rely on potentially noisy visual appearance extracted from the first frame of the video. Notably, the initialization through a skier detection algorithm has a more significant effect on the FS discipline, likely due to the initialization to a wrong skier given by the detector in multi-athlete sub-disciplines such as ski cross and dual moguls. In this condition, the tracker should be initialized by an human operator via an annotation tool to be sure it can track the target skier of interest.

Efficiency. In terms of running speed, Fig. 11 analyzes the time taken by trackers to provide skier localization at different fractions of the observed skiing performance, so to provide information on how the Delay ↓ increases while processing the video. YOLO-SORT gives the best efficiency, offering minimal delay compared to the unfolding of the skiing performance. STARK$_{SKI}$ is the least efficient tracker, it accumulates delay while processing all the video frames and finally provides the last localization of the skier over 15 s after his/her performance has ended. More in general, STARK$_{SKI}$ processes an average of 22 FPS on SkiTB. Notably, STARK$_{SKI}$ does not introduce modifications to the original STARK's neural network architecture in its STARK$_{FT-SC}$ and STARK$_{FT}$ instances. Consequently, the number of operations performed by our proposed tracker's network is, at worst (when both STARK$_{FT-SC}$ and STARK$_{FT}$ are run on the same frame), twice the number of operations executed by STARK's architecture, which is 10.9 GFLOPs (Yan et al., 2021). For comparison, STARK$_{FT}$ processes 23 FPS on SkiTB and maintains the same GFLOPs as STARK, as the neural network architecture remains unchanged. In light of this evaluation, we can state that STARK$_{SKI}$ introduces minimal additional complexity compared to the baseline STARK methodology (Yan et al., 2021).

These findings should be considered when utilizing visual trackers in computer vision-based skiing performance analytics systems. A solution based on faster tracking algorithms such as YOLO-SORT might be more suitable for systems that have to provide real-time output, such as those needed during live broadcast transmission. On the other hand, more accurate solutions as represented by STARK$_{SKI}$ might be better suited for systems where a delay in producing the analytical results is acceptable, for instance, in the production of replays during broadcasting or during video review activities conducted after the conclusion of an athlete's training session.

Video frame-rate. We evaluated the tracking performance of different tracking algorithms to understand whether the increase of video FPS can help in overcoming targets' fast motion, a characteristic of the skiing performance. SkiTB comprises videos captured at 30 FPS, aligning with broadcasting transmission conditions. Currently, we lack data with higher frame rates, preventing experiments in such conditions. However, we conducted experiments on the top four algorithms to discern the tracking accuracy trend with varying video FPS. Analyzing the trend with lower frame-rate values gives us hints on what tracking accuracy to expect with higher frame-rates. Specifically, we conducted experiments in the default configurations by considering the SkiTB's MC videos as if they were acquired at 5, 10, 15, 20, and 25 FPS. The plot in Fig. 15 visually illustrates how performance changes across these diverse settings. As evident, the STARK-based trackers (STARK$_{SKI}$, STARK$_{FT}$, STARK) exhibit minimal sensitivity to frame-rate variations. The F-Score ↑ for STARK$_{SKI}$ decreases from 0.835 at 30 FPS to 0.829 at 20 FPS, further to 0.823 at 10 FPS, and 0.817 at 5 FPS. The trend indicates a very light improvement in skier appearance localization with an increased frame-rate. While prior research (Galoogahi et al., 2017) suggests potential tracking accuracy enhancement with high-frame rate videos, we hypothesize this improvement to be marginal at 50 or 60 FPS videos, frame-rates used by some skiing performance analysis systems (Rhodin et al., 2018). Moreover, the frame-rate increase needs careful consideration in relation to the tracker speed. STARK$_{SKI}$, the best-performing algorithm, operates at an average of 22 FPS. Applying it to every frame in 60 FPS videos as for the experiments in Fig. 11 would nearly triple the processing time, diminishing the overall efficiency of the skiing performance analysis pipeline.

Conditioned fine-tuning. Table 9 presents the performance of STARK$_{FT}$ fine-tuned on the different splits available in SkiTB (and described in Section 4) which represent various real-world usage conditions. The results indicate that generalizing to unseen athletes is the most challenging application case. In other terms, this implies that the tracking algorithm tends to slightly overfit to the features of the training skier's appearance — specifically, the visual characteristics of the skiing

Fig. 12. Qualitative tracking performance. This figure shows bounding-box samples predicted by the top four trackers for frames of SkiTB's test set. STARK_{FT} and $\text{STARK}_{\text{SKI}}$ exhibit high precision in localizing both the skier's body and equipment.

Table 9
Tracking after different application-dependent fine-tuning conditions. The table reports STARK_{FT}'s F-Score ↑ under different generalization conditions. The overall result of the best application condition is highlighted in **bold**. Generalizing to unseen athletes is more challenging compared to unseen courses or newer skiing performances.

Generalization condition	AL	JP	FS	All
New performances	0.853	0.880	0.721	0.818
Unseen athletes	0.854	0.890	0.682	0.809
Unseen courses	0.861	0.917	0.808	**0.862**

suit and equipment, along with the body posture, and the motion across frames. This aspect should be considered in the deployment of fine-tuned trackers. It is advisable to explore alternative strategies (e.g., domain adaptation (Wang and Deng, 2018; Dunnhofer et al., 2021)) to prevent excessive adaptation to the athlete's appearance during training. Even though with less magnitude, generalizing to skiing performances occurring after the training ones also proves to be a demanding application condition. Consistent with our previous findings, this result indicates that fine-tuned tracking algorithms heavily rely on the visual and motion characteristics of athletes. Over different seasons, the IDs of skiers may change due to new athletes with distinct skiing styles entering professional tournaments and others retiring. Furthermore, across seasons, athletes' suits undergo changes in appearance, introducing new colors and graphics that can alter the appearance distribution, even for the same skiers. Unseen courses instead pose fewer difficulties for generalization, as the results achieved by STARK_{FT} are higher. This outcome reveals that the visual characteristics of skiing courses (i.e., gate positions, snow surface, country-specific advertising banners), along with the motion patterns of the cameras whose placement is influenced by the course's conformation, exhibit a high level of generality across various locations. Therefore, in practical application scenarios and given the diverse representation of locations in SkiTB, it is reasonable to anticipate that trackers will perform effectively in different parts of the world.

Table 10 presents the evaluation of training STARK_{FT} on multiple skiing disciplines simultaneously. Training on all the disciplines (AL, JP, FS) yields the best tracking performance. This indicates that the varied distributions of skier appearances and motion patterns across different disciplines contribute to enhancing the learning process for a more generic and generalizable representation of what a skier is and how he/she is expected to behave in videos capturing his/her performance. Notably, training and testing on separate disciplines have different impacts. FS shows the highest generalization ability to AL and JP, followed by training on AL. Indeed, FS videos encompass a higher number of different sub-disciplines (5), contributing to an increased variability in skiers' appearances and motion patterns. This overall enhancement in diversity aids in the learning process of the tracker

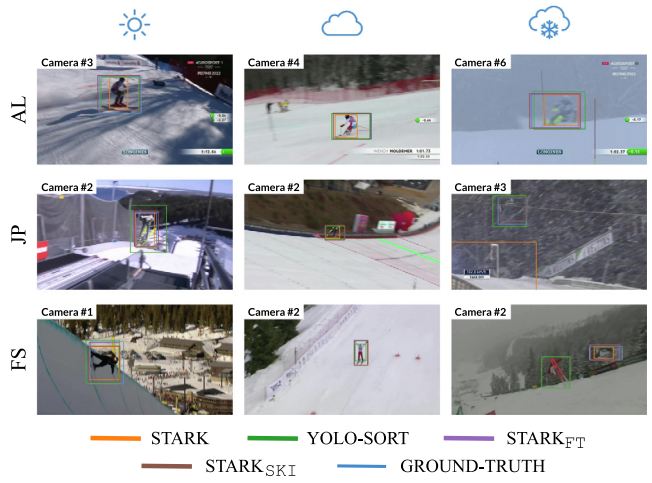


Fig. 13. Qualitative tracking performance under different weather conditions. This figure shows bounding-box samples predicted by the top four trackers for frames affected by weather conditions. Generally, all the trackers exhibit good precision in localizing both the skier's body and equipment even in the case of challenging image conditions.

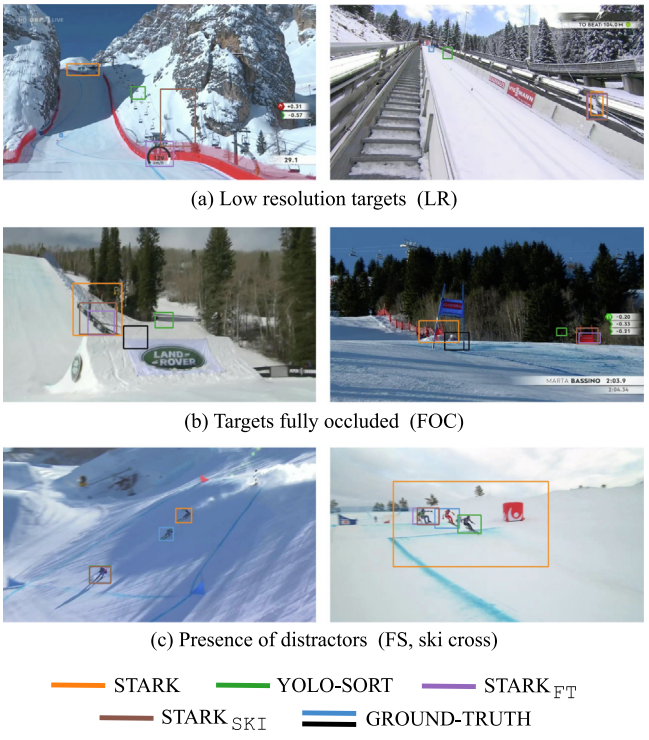


Fig. 14. Qualitative examples of tracking failure cases. This figure depicts frames on which the top trackers fail to localize the correct skier target. The first row of images shows situations where the target skier appears very small. The second row of frames in which the skier is occluded, while the third row presents frames where the target skier is surrounded by other athletes with similar appearance (□ represents the ground-truth skier's position as visible, □ as occluded).

for representing and matching skiers. Moreover, FS skiing performances are characterized by numerous jumps, which could potentially benefit the resulting motion patterns observed in ski jumping (JP). JP, with its limited diversity in courses and athlete poses, is the least helpful in improving generalization across disciplines. In fact, JP performances consist of just two sub-disciplines, where the appearance and motion of gestures exhibit limited diversity. In contrast to AL and FS gestures, the athletic gesture of ski jumping is highly repetitive. It is also executed

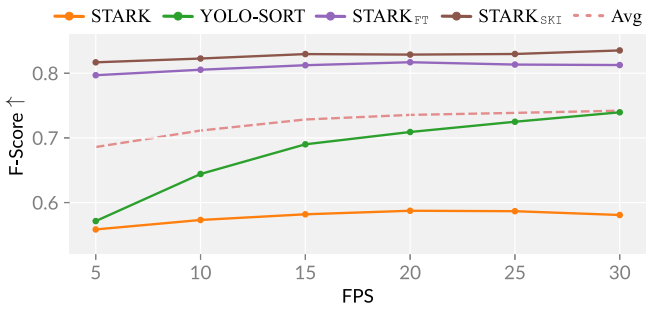


Fig. 15. Impact of video frame-rate on tracking. The plot illustrates the F-Score ↑ achieved by the top four algorithms considering SkiTB's videos at different frame-rates measured as FPS. Generally, the STARK-based trackers' tracking accuracy trend is minimally altered by reducing the number of frames per second.

Table 10
Learning tracking on different skiing disciplines. This table reports the impact of training STARK$_{FT}$ on videos capturing the different disciplines present in SkiTB. The overall result of the best set of disciplines for training is highlighted in **bold**. Indeed, training over all disciplines improves generalization generally.

Train			Test			
AL	JP	FS	AL	JP	FS	All
			0.552	0.603	0.596	0.584
✓			0.859	0.682	0.584	0.708
	✓		0.669	0.873	0.541	0.694
		✓	0.758	0.717	0.706	0.727
✓	✓	✓	0.853	0.880	0.721	**0.818**

on a standardized course, namely a ski jumping hill, which displays less variation compared to the courses found in other disciplines.

Ablation study on STARK$_{SKI}$. Table 11 presents the results of an ablation study performed over the improvements added to implement STARK$_{SKI}$. The first row presents the performance of the baseline STARK$_{FT}$. The version with Improvement 1 does not execute the code statements defined between lines 10–20 of Algorithm 1, and just executes STARK$_{FT}$ when STARK$_{FT-SC}$ is not confident. The version with Improvement 2 does not execute the code statements defined between lines 15–20, i.e. it performs the re-initialization step of STARK$_{FT-SC}$ (lines 10–12) but does not re-localize the STARK$_{FT}$ instance. The FS videos are those that benefit the most from the improvements. Lighter improvement is observed for the other disciplines, AL and JP. We also report that in the SC settings, STARK$_{FT-SC}$ achieves overall F-Score ↑, Pr ↑, Re ↑scores of 0.843, 0.834, 0.863 that are higher than the scores achieved by STARK$_{FT}$ as reported in Table 7, thus demonstrating the superiority of having a more focused search area localized around the target. It can be noticed that STARK$_{FT-SC}$ performs slightly better than STARK$_{SKI}$ in the SC setting, suggesting that sometimes the STARK$_{FT}$ instance is executed inefficiently. But executing the only STARK$_{FT-SC}$ in the MC settings results in an overall F-Score ↑ of 0.809, worse than both STARK$_{FT}$ and STARK$_{SKI}$.

Fig. 16 depicts the performance of STARK$_{SKI}$ under different δ thresholds, governing the activation of both the STARK$_{FT}$ instance (lines 8–13 of Algorithm 1) and the re-initialization of the STARK$_{FT-SC}$ instance (lines 16–20 of Algorithm 1). The overall trend, represented by the black line, highlights that a δ value of 0.5 leads to optimal tracking performance, especially in videos characterized by the FS discipline. We propose that maintaining a well-balanced value is crucial, particularly in scenarios where the STARK$_{FT}$ instance is triggered more frequently due to errors in the STARK$_{FT-SC}$ instance. For videos featuring the AL and JP disciplines, we observe that values of δ within the range [0.1, 0.9] do not significantly affect STARK$_{SKI}$'s tracking performance. This implies that the method consistently maintains tracking accuracy with high confidence values in these scenarios.

Table 11
Ablation study on the improvements introduced to implement $\text{STARK}_{\text{SKI}}$. This table presents the impact in terms of F-Score ↑ (Pr ↑/Re ↑) of the additions introduced to the baseline STARK_{FT} to develop $\text{STARK}_{\text{SKI}}$. The overall and per-discipline scores related to the best configuration are highlighted in **bold**. We observe that FS is the discipline the benefits more of the introduced components.

Version	AL	JP	FS	All
STARK_{FT}	0.853 (0.874 / 0.834)	0.880 (0.892 / 0.871)	0.721 (0.746 / 0.738)	0.818 (0.832 / 0.806)
Improvement 1	0.866 (0.885 / 0.849)	0.896 (0.898 / 0.895)	0.726 (0.731 / 0.723)	0.829 (0.838 / 0.822)
Improvement 2	0.866 (0.883 / 0.850)	0.896 (0.898 / 0.895)	0.731 (0.734 / 0.728)	0.831 (0.842 / 0.828)
$\text{STARK}_{\text{SKI}}$	**0.867** (0.885 / 0.852)	**0.896** (0.898 / 0.896)	**0.742** (0.746 / 0.738)	**0.835** (0.843 / 0.829)

Table 12
Tracking skiers' different appearance elements. The table displays STARK's F-Score ↑ (Pr ↑/Re ↑) when initialized to track target representations (bounding-box) outlining different appearance elements of a skier. The top result is highlighted in **bold**. As can be noticed, the score achieved by tracking the appearance of both skier's body and equipment is very close to the one achieved by tracking the body only. Tracking the appearance of the skis instead results more difficult.

Target bounding-box representation	F-Score ↑ (Pr ↑/Re ↑)
Skier's body and equipment	0.751 (0.751/0.751)
Skier's body only	**0.759** (0.759/0.759)
Skier's skis only	0.512 (0.511/0.511)

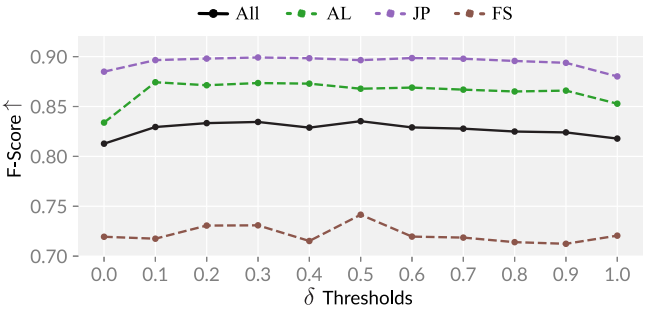


Fig. 16. Impact of δ threshold values on $\text{STARK}_{\text{SKI}}$. This plot depict the F-Score ↑ values achieved by $\text{STARK}_{\text{SKI}}$ on the different skiing discipline when different values of δ are applied to the confidence c_t predicted by $\text{STARK}_{\text{FT-SC}}$ and STARK_{FT} inside $\text{STARK}_{\text{SKI}}$. Generally, $\delta = 0.5$ results to the highest tracking accuracy due to the clear advantage brought on the FS videos. For AL and JP, changing the values in the range $[0.1, 0.9]$ leads to approximately the same tracking performance.

We conducted experiments to assess the generalization of the $\text{STARK}_{\text{SKI}}$ pipeline to other tracker instances. For this purpose, we introduced $\text{MixFormer}_{\text{SKI}}$, comprising two instances, $\text{MixFormer}_{\text{FT-SC}}$ and $\text{MixFormer}_{\text{FT}}$, both implemented and fine-tuned with the same configuration hyper-parameters as STARK_{FT} and $\text{STARK}_{\text{FT-SC}}$. In the MC experimental setting, $\text{MixFormer}_{\text{SKI}}$ achieves an overall F-Score ↑ of 0.836, surpassing the baseline $\text{MixFormer}_{\text{FT}}$ score of 0.829. These findings emphasize the importance of developing algorithms tailored to the specific scenarios present in broadcasting skiing videos with consistent skier appearance motions followed by abrupt and sudden frame changes.

Target representation. We performed experiments to understand the influence of background appearance within the initialization box that in SkiTB outlines both the skier's body and equipment. The composition of the human body and the skis form a particular appearance (like a vertically reversed T, see the first frame of the 4th row in Fig. 2) that, when outlined with axis-aligned bounding-boxes results in much background information be included in the resulting image patch. Hence, we evaluated the tracking accuracy when the box delineating solely the body's appearance is used. By leveraging the key-point annotations for distinct human body parts, skis, and poles available in the Ski2DPose dataset (Bachmann et al., 2019) we obtained tracks for such appearance elements by considering the bounding-box that encloses their specific appearance, in the same fashion as described in Section 6.2. We could not execute this evaluation with SkiTB because it currently does not provide bounding-box nor pose key-points for such elements. Therefore, on such sparsely annotated tracks the OPE protocol and F-Score ↑ measure were employed to evaluate the best generic object tracker STARK. The outcomes, presented in Table 12, substantiate that the tracker encounters no pronounced difficulties when tracking the combined appearance of the body and equipment, relative to exclusively tracking the body's appearance. Conversely, tracking the equipment in isolation (e.g., the athlete's skis) presents a considerably more challenging task.

7.3. Impact on applications

Finally, Table 13 presents the impact of the trackers on the 2D skier pose estimation tasks described in Section 6.2. Generally, we observe that employing skier-specific trackers (YOLO-SORT, STARK_{FT}, $\text{STARK}_{\text{SKI}}$) improves the skier tracking results as well as the pose estimation results. For AL on Ski2DPose, $\text{STARK}_{\text{SKI}}$ and STARK_{FT} have nearly the same impact on AlphaPose. They improve the pose estimation based on the generic object STARK by 20% in PCK ↑ and by 44% in MPJPE ↓. $\text{STARK}_{\text{SKI}}$ and YOLO-SORT have comparable impact in JP as represented by the YouTube Skijump dataset. With respect to STARK, they improve the PCK ↑ and MPJPE ↓ scores by 15% and ∼20% respectively. Across disciplines, $\text{STARK}_{\text{SKI}}$ results in the best generalizing tracker by tracking and impact scores. Overall, these results show that the trackers' performances on SkiTB reflect the impact on high-level skiing understanding tasks. Thus, we expect $\text{STARK}_{\text{SKI}}$ and other skier-specific tracking methodologies to be beneficial as instance-specific athlete appearance localizers in the development of computer vision-systems for skiing performance analysis. It is worth mentioning that these results are obtained with the limited annotations present in the respective small-scale datasets. We hypothesize the relation with the SkiTB's results to become more evident on more densely-labeled datasets.

8. Conclusions

This paper presented a comprehensive study on tracking skiing athletes in monocular multi-camera broadcasting videos. Through the evaluation of established and newly introduced methodologies on the newly released dataset SkiTB, the study revealed that fine-tuned application-specific deep learning-based algorithms demonstrate consistent tracking performance and promising applicability throughout a skier's performance. These trackers exhibit robustness under various conditions such as challenging weather, fast camera motion, scale changes, and background clutter, and they generalize well to new locations of application. However, the study also identified certain limitations that prevent the methods to be perfect. Challenges arise in maintaining a continuous per-frame reference to the target skier across camera shotcuts, in accurately localizing the skier in the presence of distractors, small appearance, occlusion, and fast motion. Additionally, the generalization to unseen athletes poses particular difficulties. Top-performance trackers should be also improved in their efficiency. Regarding generic

Table 13
Impact of trackers on high-level skiing performance understanding tasks. We report the impact of the top trackers' predictions in the task of 2D body and equipment pose estimation for the alpine skiing (AL) and ski jumping (JP) disciplines. The Tracking and Pose Estimation results related to the best tracking algorithm are highlighted in **bold**. As can be expected, more accurate trackers lead to a more accurate pose prediction in general. (GT boxes were extracted from the annotated pose key-points).

Discipline	Dataset	Task	Metric	STARK	YOLO-SORT	STARK_{FT}	$\text{STARK}_{\text{SKI}}$	GT box
AL	Ski2DPose	Tracking	F-Score ↑	0.751	0.830	0.848	**0.849**	–
		Pose Estimation	PCK ↑/MPJPE ↓	0.573/0.059	0.685/0.034	**0.694**/ **0.033**	0.686/0.033	0.682/0.036
JP	YouTube Skijump	Tracking	F-Score ↑	0.670	0.748	0.768	**0.775**	–
		Pose Estimation	PCK ↑/MPJPE ↓	0.516/0.029	**0.598**/**0.026**	0.574/0.023	0.596/0.026	0.571/0.026

object trackers, we observed that they struggle to generalize to the domain of interest, demonstrating that they are still not able to generalize as well as humans to situations different from those observed in training sets.

Future work should focus on addressing these limitations. Solutions may involve refining skier-specific tracking methods, improving their generalization, and developing strategies to better integrate with high-level skiing performance understanding modules. Furthermore, the SkiTB will be extended with new annotations (e.g. human poses) to enable the research and development of more sophisticated and effective skiing performance analysis tools.

CRediT authorship contribution statement

Matteo Dunnhofer: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Christian Micheloni:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

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.

Data availability

The data used will be made public upon the acceptance of the paper.

Acknowledgments

Research supported by the project between the University of Udine and the organizing committee of EYOF 2023 Friuli-Venezia Giulia.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT (GPT 3.5) in order to enhance the clarity and readability of part of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Bachmann, R., Spörri, J., Fua, P., Rhodin, H., 2019. Motion capture from pan-tilt cameras with unknown orientation. In: International Conference on 3D Vision. 3DV.
Bertasius, G., Soo Park, H., Yu, S.X., Shi, J., 2017. Am I a baller? basketball performance assessment from first-person videos. In: CVPR.
Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H., 2016. Fully-convolutional siamese networks for object tracking. In: ECCVW.
Bettadapura, V., Pantofaru, C., Essa, I., 2016. Leveraging contextual cues for generating basketball highlights. In: ACM MM.
Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. In: ICIP.
Bhat, G., Danelljan, M., Van Gool, L., Timofte, R., 2019. Learning discriminative model prediction for tracking. In: ICCV.
Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M., 2010. Visual object tracking using adaptive correlation filters. In: CVPR.
Borsuk, V., Vei, R., Kupyn, O., Martyniuk, T., Krashenyi, I., Matas, J., 2022. FEAR: Fast, efficient, accurate and robust visual tracker. In: ECCV.
Cai, Y., Liu, J., Tang, J., Wu, G., 2023. Robust object modeling for visual tracking. In: ICCV.
Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR.
Čehovin, L., Kristan, M., Leonardis, A., 2013. Robust visual tracking using an adaptive coupled-layer visual model. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2012.145.
Chappa, N.V.R., Nguyen, P., Nelson, A.H., Seo, H.-S., Li, X., Dobbs, P.D., Luu, K., 2023. SPARTAN: Self-supervised spatiotemporal transformers approach to group activity recognition. In: CVPRW.
Chen, X., Peng, H., Wang, D., Lu, H., Hu, H., 2023. SeqTrack: Sequence to sequence learning for visual object tracking. In: CVPR.
Chen, F., Wang, X., Zhao, Y., Lv, S., Niu, X., 2022. Visual object tracking: A survey. Comput. Vis. Image Underst.
Cheng, B., Li, J., Chen, Y., Zeng, T., 2023. Snow mask guided adaptive residual network for image snow removal. Comput. Vis. Image Underst.
Choi, J., Kwon, J., Lee, K.M., 2018. Real-time visual tracking by deep reinforced decision making. Comput. Vis. Image Underst.
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: CVPRW.
Comaniciu, D., Ramesh, V., Meer, P., 2000. Real-time tracking of non-rigid objects using mean shift. In: CVPR. http://dx.doi.org/10.1109/CVPR.2000.854761.
Cui, Y., Jiang, C., Wang, L., Wu, G., 2022. Mixformer: End-to-end tracking with iterative mixed attention. In: CVPR.
Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L., 2023. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. arXiv preprint arXiv:2304.05170.
Dai, K., Zhang, Y., Wang, D., Li, J., Lu, H., Yang, X., 2020. High-performance long-term tracking with meta-updater. In: CVPR.
Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., 2019. ATOM: Accurate tracking by overlap maximization. In: CVPR.
Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Leal-Taixé, L., 2021. Motchallenge: A benchmark for single-camera multiple target tracking. Int. J. Comput. Vis.
Drory, A., Zhu, G., Li, H., Hartley, R., 2017. Automated detection and tracking of slalom paddlers from broadcast image sequences using cascade classifiers and discriminative correlation filters. Comput. Vis. Image Underst.
Dunnhofer, M., Furnari, A., Farinella, G.M., Micheloni, C., 2023a. Visual object tracking in first person vision. Int. J. Comput. Vis.
Dunnhofer, M., Martinel, N., Luca Foresti, G., Micheloni, C., 2019. Visual tracking by means of deep reinforcement learning and an expert demonstrator. In: ICCVW.
Dunnhofer, M., Martinel, N., Micheloni, C., 2020. Tracking-by-trackers with a distilled and reinforced model. In: ACCV.
Dunnhofer, M., Martinel, N., Micheloni, C., 2021. Weakly-supervised domain adaptation of deep regression trackers via reinforced knowledge distillation. IEEE RA-L.
Dunnhofer, M., Micheloni, C., 2022. CoCoLoT: Combining complementary trackers in long-term visual tracking. In: ICPR.
Dunnhofer, M., Simonato, K., Micheloni, C., 2022. Combining complementary trackers for enhanced long-term visual object tracking. Image Vis. Comput.
Dunnhofer, M., Sordi, L., Micheloni, C., 2023b. Visualizing skiers' trajectories in monocular videos. In: CVPRW.
Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H., 2019. LaSOT: A high-quality benchmark for large-scale single object tracking. In: CVPR.
Fan, H., Miththanthaya, H.A., Harshit, Rajan, S.R., Liu, X., Zou, Z., Lin, Y., Ling, H., 2021. Transparent object tracking benchmark. In: ICCV.
Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., Lu, C., 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans. Pattern Anal. Mach. Intell.
Feng, N., Song, Z., Yu, J., Chen, Y.P.P., Zhao, Y., He, Y., Guan, T., 2020. SSET: a dataset for shot segmentation, event detection, player tracking in soccer videos. Multimedia Tools Appl. 28971–28992.
Gadde, C.A., Jawahar, C., 2022. Transductive weakly-supervised player detection using soccer broadcast videos. In: WACV.
Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S., 2017. Need for speed: A benchmark for higher frame rate object tracking. In: ICCV.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.
Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.M., Hicks, S.L., Torr, P.H., 2016. Struck: Structured output tracking with kernels. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2015.2509974.
Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision. Cambridge University Press.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR.
Held, D., Thrun, S., Savarese, S., 2016. Learning to track at 100 FPS with deep regression networks. In: ECCV.
Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell.
Honda, Y., Kawakami, R., Yoshihashi, R., Kato, K., Naemura, T., 2022. Pass receiver prediction in soccer using video and players' trajectories. In: CVPRW.
Hu, S., Zhao, X., Huang, L., Huang, K., 2023. Global instance tracking: Locating target more like humans. IEEE Trans. Pattern Anal. Mach. Intell.
Huang, L., Zhao, X., Huang, K., 2019. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Trans. Pattern Anal. Mach. Intell.
Huang, L., Zhao, X., Huang, K., 2020. GlobalTrack: A simple and strong baseline for long-term tracking. In: AAAI.
International Olympic Committee, 2023. History of alpine skiing. URL: https://olympics.com/en/sports/alpine-skiing/.
International Ski and Snowboard Federation, URL: https://www.fis-ski.com.
Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, Rai, P., 2020. Ultralytics/yolov5: v3.1 - Bug fixes and performance improvements. http://dx.doi.org/10.5281/zenodo.4154370.
Koshkina, M., Pidaparthy, H., Elder, J.H., 2021. Contrastive learning for sports video: Unsupervised player classification. In: CVPRW.
Kou, Y., Gao, J., Li, B., Wang, G., Hu, W., Wang, Y., Li, L., 2023. ZoomTrack: Target-aware non-uniform resizing for efficient visual tracking. In: NeurIPS.
Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Chang, H.J., Danelljan, M., Zajc, L.Č., Lukežič, A., Drbohlav, O., et al., 2023. The tenth visual object tracking VOT2022 challenge results. In: ECCVW.
Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Danelljan, M., Zajc, L.Č., Lukežič, A., Drbohlav, O., He, L., Zhang, Y., Yan, S., Yang, J., Fernández, G., et al., 2020. The eighth visual object tracking VOT2020 challenge results. In: Bartoli, A., Fusiello, A. (Eds.), ECCVW.
Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kämäräinen, J.-K., Chang, H.J., Danelljan, M., Cehovin, L., Lukežič, A., Drbohlav, O., Käpylä, J., Häger, G., Yan, S., Yang, J., Zhang, Z., Fernández, G., 2021. The ninth visual object tracking VOT2021 challenge results. In: ICCVW.
Kristan, M., Perš, J., Perše, M., Kovačič, S., 2009. Closed-world tracking of multiple interacting targets for indoor-sports applications. Comput. Vis. Image Underst.
Li, X., Chuah, M.C., 2018. Rehar: Robust and efficient human activity recognition. In: WACV.
Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2019. SIAMRPN++: Evolution of siamese visual tracking with very deep networks. In: CVPR.
Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H., 2022. Swintrack: A simple and strong baseline for transformer tracking. In: NeurIPS.
Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: ECCV.
Liu, J., Carr, P., Collins, R.T., Liu, Y., 2013. Tracking sports players with context-conditioned motion models. In: CVPR.
Lu, J., Li, S., Guo, W., Zhao, M., Yang, J., Liu, Y., Zhou, Z., 2023. Siamese graph attention networks for robust visual object tracking. Comput. Vis. Image Underst.
Ludwig, K., Harzig, P., Lienhart, R., 2022. Detecting arbitrary intermediate keypoints for human pose estimation with vision transformers. In: WACVW.
Ludwig, K., Kienzle, D., Lorenz, J., Lienhart, R., 2023. Detecting arbitrary keypoints on limbs and skis with sparse partly correct segmentation masks. In: WACVW.
Lukeźič, A., Zajc, L.Č., Vojíř, T., Matas, J., Kristan, M., 2020. Performance evaluation methodology for long-term single-object tracking. IEEE Trans. Cybern.
Maresca, M.E., Petrosino, A., 2013. MATRIOSKA: A multi-level approach to fast tracking by learning. In: International Conference on Image Analysis and Processing. ICIAP.
Matsumura, S., Mikami, D., Saijo, N., Kashino, M., 2021. Spatiotemporal motion synchronization for snowboard big air. In: 1st Workshop on Computer Vision for Winter Sports at WACV 2022.
Mauthner, T., Koch, C., Tilp, M., Bischof, H., 2007. Visual tracking of athletes in beach volleyball using a single camera. Int. J. Comput. Sci. Sport.
Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D.P., Yu, F., Van Gool, L., 2022. Transforming model prediction for tracking. In: CVPR.
Mayer, C., Danelljan, M., Paudel, D.P., Gool, L.V., 2021. Learning target candidate association to keep track of what not to track. In: ICCV.
Morimitsu, H., Bloch, I., Cesar-Jr, R.M., 2017. Exploring structure for long-term tracking of multiple objects in sports videos. Comput. Vis. Image Underst.
Mueller, M., Smith, N., Ghanem, B., 2016. A benchmark and simulator for UAV tracking. In: ECCV.
Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B., 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In: ECCV. Springer Verlag.
Nam, H., Han, B., 2016. Learning multi-domain convolutional neural networks for visual tracking. In: CVPR.
Pidaparthy, H., Dowling, M.H., Elder, J.H., 2021. Automatic play segmentation of hockey videos. In: CVPRW.
Pidaparthy, H., Elder, J., 2019. Keep your eye on the puck: Automatic hockey videography. In: WACV.
Qi, J., Li, D., Zhang, C., Wang, Y., 2022. Alpine skiing tracking method based on deep learning and correlation filter. IEEE Access.
Quiroga, J., Carrillo, H., Maldonado, E., Ruiz, J., Zapata, L.M., 2020. As seen on TV: Automatic basketball video production using Gaussian-based actionness and game states recognition. In: CVPRW.
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: CVPR.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS.
Rhodin, H., Meyer, F., Spörri, J., Müller, E., Constantin, V., Fua, P., Katircioglu, I., Salzmann, M., 2018. Learning monocular 3D human pose estimation from multi-view images. In: CVPR. http://dx.doi.org/10.1109/CVPR.2018.00880.
Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV.
Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., TOsmanov, Kruchinin, D., Zankevich, A., DmitriySidnev, Markelov, M., Johannes222, Chenuet, M., a-andre, telenachos, Melnikov, A., Kim, J., Ilouz, L., Glazov, N., Priya4607, Tehrani, R., Jeong, S., Skubriev, V., Yonekura, S., vugia truong, zliang7, lizhming, Truong, T., 2020. Opencv/cvat: v1.1.0. http://dx.doi.org/10.5281/zenodo.4009388.
Steinkellner, P., Schöffmann, K., 2021. Evaluation of object detection systems and video tracking in skiing videos. In: 2021 International Conference on Content-Based Multimedia Indexing. CBMI, http://dx.doi.org/10.1109/CBMI50038.2021.9461905.
Štepec, D., Skočaj, D., 2022. Video-based ski jump style scoring from pose trajectory. In: WACVW.
The Nielsen Company, 2022a. FIS alpine skiing world cup report 2021-22. URL: https://assets.fis-ski.com/image/upload/v1653461303/fis-prod/assets/FIS_Alpine_Skiing_World_Cup_Report_2021-22_short.pdf.
The Nielsen Company, 2022b. FIS freestyle ski world cup report 2021-22. URL: https://assets.fis-ski.com/image/upload/v1653461304/fis-prod/assets/FIS_Freestyle_Ski_World_Cup_Report_2021_22_short.pdf.
The Nielsen Company, 2022c. Viessmann FIS ski jumping world cup men 2021–2022. URL: https://assets.fis-ski.com/image/upload/v1653461308/fis-prod/assets/Viessmann_FIS_Ski_Jumping_World_Cup_Men_2022_short.pdf.
Theiner, J., Ewerth, R., 2023. TVCalib: Camera calibration for sports field registration in soccer. In: WACV.
Theiner, J., Gritz, W., Müller-Budack, E., Rein, R., Memmert, D., Ewerth, R., 2022. Extraction of positional player data from broadcast soccer videos. In: WACV.
Thomas, G., Gade, R., Moeslund, T.B., Carr, P., Hilton, A., 2017. Computer vision for sports: Current applications and research topics. Comput. Vis. Image Underst.
Vanat, L., 2022. 2022 International report on snow & mountain tourism. URL: https://www.vanat.ch/RM-world-report-2022.pdf.
Vandeghen, R., Cioppa, A., Van Droogenbroeck, M., 2022. Semi-supervised training to improve player and ball detection in soccer. In: CVPRW.
Vats, K., Fani, M., Clausi, D.A., Zelek, J., 2021. Puck localization and multi-task event recognition in broadcast hockey videos. In: CVPRW.
Vats, K., McNally, W., Walters, P., Clausi, D.A., Zelek, J.S., 2022. Ice hockey player identification via transformers and weakly supervised learning. In: CVPRW.
Wandt, B., Rudolph, M., Zell, P., Rhodin, H., Rosenhahn, B., 2021. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In: CVPR.
Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neurocomputing 312.
Wang, J., Qiu, K., Peng, H., Fu, J., Zhu, J., 2019. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In: ACM MM.
Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y., 2023. Autoregressive visual tracking. In: CVPR.
Wu, Y., Lim, J., Yang, M.-H., 2015. Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell.
Yan, B., Jiang, Y., Sun, P., Wang, D., Yuan, Z., Luo, P., Lu, H., 2022. Towards grand unification of object tracking. In: ECCV.
Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021. Learning spatio-temporal transformer for visual tracking. In: ICCV.
Yan, B., Zhao, H., Wang, D., Lu, H., Yang, X., 2019. 'Skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In: ICCV.
Ye, B., Chang, H., Ma, B., Shan, S., Chen, X., 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In: ECCV.
Yun, S., Choi, J., Yoo, Y., Yun, K., Choi, J.Y., 2017. Action-decision networks for visual tracking with deep reinforcement learning. In: CVPR.
Zhu, Y., Yan, W.Q., 2022. Ski fall detection from digital images using deep learning. In: International Conference on Control and Computer Vision.
Zwölfer, M., Heinrich, D., Schindelwig, K., Wandt, B., Rhodin, H., Spoerri, J., Nachbauer, W., 2021. Improved 2D keypoint detection in out-of-balance and fall situations – combining input rotations and a kinematic model. In: 1st Workshop on Computer Vision for Winter Sports at WACV 2022.