Tracking Skiers from the Top to the Bottom

Matteo Dunnhofer Luca Sordi Niki Martinel Christian Micheloni
Machine Learning and Perception Lab, University of Udine, Udine, Italy

Abstract

Skiing is a popular winter sport discipline with a long history of competitive events. In this domain, computer vision has the potential to enhance the understanding of athletes' performance, but its application lags behind other sports due to limited studies and datasets. This paper makes a step forward in filling such gaps. A thorough investigation is performed on the task of skier tracking in a video capturing his/her complete performance. Obtaining continuous and accurate skier localization is preemptive for further higher-level performance analyses. To enable the study, the largest and most annotated dataset for computer vision in skiing, SkiTB, is introduced. Several visual object tracking algorithms, including both established methodologies and a newly introduced skier-optimized baseline algorithm, are tested using the dataset. The results provide valuable insights into the applicability of different tracking methods for vision-based skiing analysis. SkiTB, code, and results are available at `https://machinelearning.uniud.it/datasets/skitb`.

1. Introduction

Skiing is a globally recognized and highly popular winter sport discipline [74], with a long history of competitive events dating back to the 1840s [15]. It has been a significant part of the Winter Olympic Games since their inception in 1924 [15]. Today, skiing comprehends various disciplines such as alpine skiing, ski jumping, and freestyle skiing. [70]. Their competitive form holds a prominent position within the winter sports industry, generating over 1.7 billion media views during a winter season [16–18].

Applying data-driven analytics to the skiing performance can improve the technical ability and physical well-being of athletes and augment the educational value and entertainment of broadcasting recordings of professional competitions, overall resulting in more impressive, safe, and engaging competitions. In such applications, computer vision offers promising opportunities for capturing and analyzing skiing performances without relying on wearable sensors. However, the use of computer vision in skiing

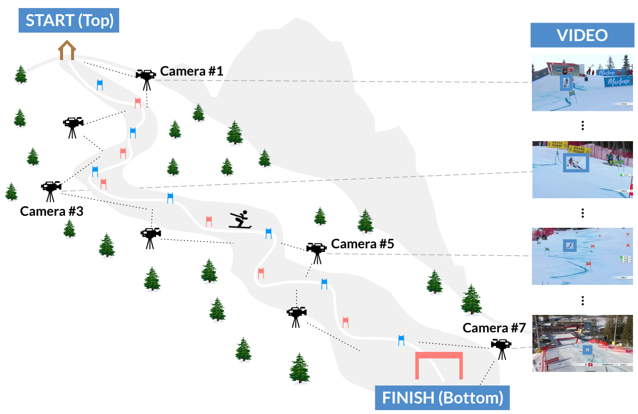

Figure 1. **Tracking a skier from the top to the bottom of the course.** This paper focuses on applying visual object tracking algorithms to localize a skier per-frame (e.g. with bounding-boxes □) in a video capturing his/her complete performance. Due to the large spatial extent of skiing courses, multiple 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.

has been limited compared to other sports such as soccer [13, 33, 40, 72, 73, 75], basketball [2, 4, 11, 50, 65], or ice-hockey [45, 62, 63, 76, 77].

Previous research in this domain has primarily focused on reconstructing skiers' poses in 2D or 3D [1, 52] and understanding the style of ski jumps [78, 80]. A crucial step in all these approaches involves localizing the skier within the video frames, as this is essential for the subsequent higher-level computational analysis. The methods usually rely on off-the-shelf or fine-tuned object detection models [51, 66, 67] without utilizing the temporal information available in the athlete's performance evolution. 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 present several unique challenges characterized by exercises performed with unique body-equipment relations, at high speed, and on a continuously changing playing field subject to extreme outdoor winter weather conditions. All of these raise important questions regarding their influence on image-based systems, and addressing them in a systematic and extensive manner could have implications even for the computer vision community as a whole, as evidenced by recent activities.[1]

This paper focuses on the study of tracking a skier in monocular broadcasting videos of professional skiing competitions. A new visual tracking dataset named SkiTB ("Skiers from the Top to the Bottom") is introduced to address the lack of appropriate benchmarks. It consists of 300 video recordings across the most challenging skiing disciplines. The videos cover the skiers' complete performance, from the top to the bottom of the course, as exemplified by Figure 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, 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.

In sum, the contributions of this paper are:

- A systematic and in-depth analysis of skier tracking in videos, which has not been thoroughly studied in previous works.
- The description and release of SkiTB a novel benchmark dataset curated specifically for evaluating computer vision-based systems in skiing. The dataset is designed to be diverse, representative, and densely labeled.
- $\text{STARK}_{\texttt{SKI}}$, a baseline algorithm optimized for skier tracking in multi-camera videos.

2. Related Work

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. Early trackers utilized mean shift algorithms [14], key-point [55], part-based techniques [10], or SVM learning [36]. Correlation filters gained popularity due to their fast processing [7, 39]. More recently, deep learning-based solutions, including regression networks [26, 38], online tracking-by-detection methods [61], reinforcement learning-based methods [25, 85], deep discriminative correlation filters [6, 21], siamese network-based trackers [3, 49], and transformers [19, 57, 82, 84], led to higher tracking accuracy. Long-term trackers and methods combining multiple trackers have been also explored [27, 43, 83].

Such a progress in SOT algorithms is attributed to well-curated evaluation datasets featuring diverse object types [46–48, 81] and large-scale datasets for visual object tracking in generic domains [30, 42, 60]. Application-centric benchmarks exist for specific domains such as drones [59], high frame-rate videos [34], transparent objects [31], and egocentric videos [24]. These benchmarks contribute to the development of accurate and reliable tracking systems in specific application scenarios.

The aforementioned datasets [30, 42, 46, 60, 81] 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.

Applications of Computer Vision to Skiing. Recent advancements in computer vision [9, 37, 67] have enabled vision-based applications in skiing performance analysis. For example, [86] proposed object detection and human pose estimation algorithms to recognize falls of alpine skiers, while [87] discussed the combination of pose estimation with kinematics models. [1] introduced a methodology to reconstruct 3D poses from images captured by multiple cameras observing a single slope section. Ski jumping analysis involved scoring the style of jumps using 2D human pose trajectories [78] and detecting key-points on the human body and skis in still images using improved vision transformer architectures [52, 53]. For freestyle skiing and snowboarding, algorithms were developed to evaluate the quality of jumps in monocular videos [80] and to synchronize videos for comparing the timing and spatial extent of aerial maneuvers [56].

The discussed pipelines present object detection [66, 67] or off-the-shelf visual tracking [80] 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 [64, 71]. These studies fo-

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

Table 1. **Comparison of SkiTB with publicly available skiing-related datasets.** This table shows a comparison between some key statistics of our SkiTB and of other datasets for computer vision tasks available in the skiing domain. As can be noticed, ours results in the largest, most diverse, and most annotated dataset. (n/a stands for "not annotated".)

Dataset	Skimovie [71]	Ski2DPose [1]	SkiPosePTZ [1, 68]	YouTube Skijump [53]	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	352978
# 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

cused 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 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 the latter 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 skier's position and size 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 [66, 67] 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 [46, 54], specifically of the global instance variation [41] which aims to continuously localize a target object over an extended period, even across camera shot-cuts. Figure 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 localization and does not cover the broader task of tracking a skier across a network of automatically controlled cameras. We believe that our findings can contribute to the development of such technology, since to control cameras the athlete must be first localized in the frames of the video stream.

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. Particularly, we adhere to the following design principles:[2]

- 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.

[2]Further details are given in Appendix A of the supp. document.

Table 2. **Key statistics of SkiTB.** The following table offers overall and per-skiing discipline information about the 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	215517	38201	99260	352978
# 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

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.

Video Collection. The videos in SkiTB were 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 International Ski and Snowboard Federation (FIS) [70]: 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 [1,53,71] did not encompass all the desired characteristics, necessitating the creation of a new video collection. The videos have a framerate of 25/30 FPS and resolutions ranging from 360p to 720p. More detailed statistics are in Table 2.

Frame-level Annotations. Each of the frames belonging to the 300 multi-camera (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 Figure 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 on 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

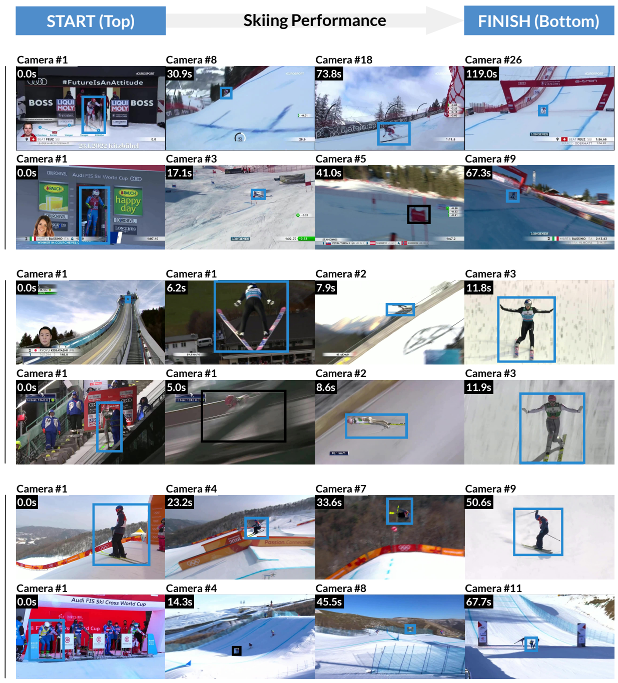

Figure 2. **Frame and bounding-box samples from SkiTB.** We showcase examples of video frames from our dataset for the different disciplines: alpine skiing (AL), ski jumping (JP), and freestyle skiing (FS). Each frame is accompanied by a manually annotated bounding-box. A blue rectangle (□) localizes 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.

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 for employing bounding-boxes as target representations stems from the requirement, by higher-level performance understanding modules (e.g. 2D or 3D pose detectors [1,53,78]), of a rectangle-shaped localization to extract athlete-specific visual information. In this view, bounding-boxes offer a simpler yet sufficiently informative representation that provides the position and the size of the athlete's appearance in the frames. 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.

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 subsequences 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's 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. Following the established practice of visual object tracking benchmarks [24,30,46,59,81], 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. All of these labels can be utilized for video clustering, enabling experimental results to be conditioned on different characteristics of the domain.

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 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.

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 [69] 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 [70].

5. Trackers

Trackers for Generic Objects. 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 [54], as in our application of interest. The trackers falling in this category include SPLT [83], GlobalTrack [43], LTMU [20], KeepTrack [58], STARK [82], and CoCoLoT [27, 28]. In addition, we included the short-term trackers [54] MOSSE [7], KCF [39], and SiamRPN++ [49] for their general popularity, and MixFormer [19], OSTrack [84], FEAR [8], and SeqTrack [12] 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 [30] and could be even suitable for skier localization in SC tracking tasks.

Skier-specific Trackers. We also assessed the performance of baseline trackers specifically designed for skier tracking. The YOLO-SORT tracker adopts a tracking-by-detection approach utilizing an YOLOX instance [35, 44] (fine-tuned on SkiTB's training set) and the SORT [5] algorithm to detect and associate boxes across consecutive frames. $\text{STARK}_{\texttt{FT}}$ instead implements an instance of STARK (STARK-ST50) [82] fine-tuned on SkiTB using the original hyper-parameter values.

$STARK_{SKI}$. Furthermore, we introduce a new tracker, called $\text{STARK}_{\texttt{SKI}}$, which consists of two instances of $\text{STARK}_{\texttt{FT}}$. The first instance serves as a precise skier tracker during periods of target visibility, typically within an SC clip where skier and camera move smoothly. The second instance acts as a skier re-detector, activated when the confidence c_t of the first instance is low, often occurring after camera shot-cut or target occlusion. The first instance uses a smaller factor (3.0 instead of 5.0) to compute a search area with higher resolution that includes more information of the target's visual appearance, resulting in more accurate bounding-box predictions. The second instance re-locates the target whenever the confidence of the first instance drops below 0.5. It uses the standard search area factor (5.0) to find the skier in a larger frame area centered at the position predicted by the first instance. Pseudo-code and further details are given in Appendix B of the supp. document.

6. Evaluation

6.1. Tracking Accuracy

Evaluation Protocol. We utilized the one-pass evaluation (OPE) protocol [81] to run trackers for performance assessment. This protocol closely simulates tracking implementation in real application conditions. It involves initializing the tracker with a target bounding-box in the first frame and running it on subsequent frames until the end of the video. In our case, this means initializing the tracker at the beginning of the skier's performance and running it until the end. By default, our experiments were conducted using the ground-truth bounding-box for initialization, thus providing the best possible evaluation conditions for the trackers. By considering practical deployment scenarios where automatic athlete localization is required (such as real-time analysis during broadcasting), we also tested the trackers by initialization with the box predicted by the YOLOX de-

Table 3. **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. In general, 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	MixFormer	KeepTrack	OSTrack	SeqTrack	LTMU	CoCoLoT	STARK	YOLO-SORT	STARK_{FT}	$\text{STARK}_{\text{SKI}}$
All	0.093 0.092 0.094	0.061 0.061 0.062	0.248 0.270 0.235	0.338 0.419 0.301	0.493 0.493 0.495	0.526 0.518 0.535	0.527 0.555 0.508	0.528 0.520 0.537	0.534 0.538 0.533	0.554 0.565 0.545	0.562 0.572 0.555	0.584 0.595 0.576	0.740 ③ 0.730 0.751	0.818 ② 0.832 0.806	0.835 ① 0.843 0.829
AL	0.031 0.031 0.032	0.024 0.024 0.024	0.144 0.143 0.145	0.270 0.260 0.229	0.485 0.487 0.483	0.463 0.458 0.468	0.518 0.561 0.484	0.462 0.457 0.467	0.479 0.485 0.475	0.524 0.541 0.509	0.532 0.546 0.521	0.552 0.565 0.540	0.798 0.790 0.807	0.853 0.874 0.834	0.868 0.885 0.852
JP	0.155 0.153 0.157	0.098 0.097 0.099	0.281 0.310 0.262	0.373 0.451 0.338	0.504 0.507 0.502	0.574 0.567 0.581	0.536 0.576 0.510	0.577 0.571 0.584	0.590 0.598 0.584	0.576 0.591 0.565	0.584 0.606 0.569	0.603 0.630 0.582	0.818 0.807 0.830	0.880 0.892 0.871	0.896 0.898 0.896
FS	0.092 0.090 0.094	0.065 0.067 0.080	0.319 0.358 0.298	0.372 0.446 0.336	0.491 0.483 0.500	0.541 0.529 0.556	0.528 0.528 0.530	0.545 0.532 0.560	0.533 0.530 0.539	0.562 0.564 0.562	0.570 0.564 0.577	0.596 0.590 0.604	0.603 0.592 0.616	0.721 0.730 0.713	0.742 0.746 0.738

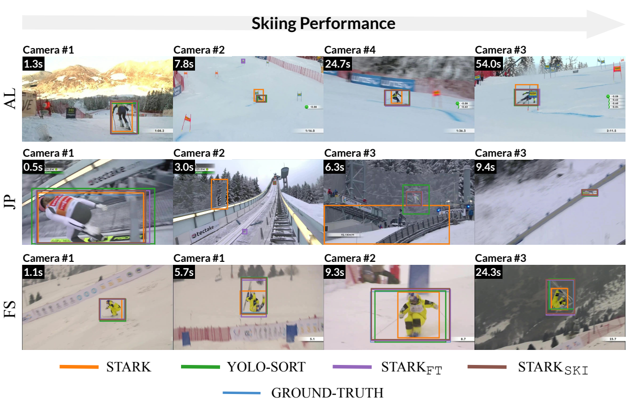

Figure 3. **Qualitative tracking performance.** This figure shows bounding-box samples predicted by the top four trackers for frames of SkiTB's test set. STARK_{FT} and $\text{STARK}_{\text{SKI}}$ exhibit high precision in localizing both the skier's body and equipment.

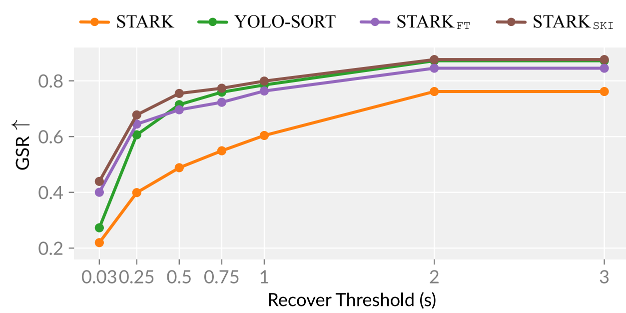

Figure 4. **Fraction of consistent skier tracking starting from the top.** This plot depicts the average fraction of consecutive frames in which the target skier is accurately localized before losing track, measured as the GSR score [24]. Various time thresholds in seconds are employed to assess the trackers' ability to recover from failures over time [47].

tector [44, 66] trained on SkiTB's training set. If not specified otherwise, in the experiments all the trackers were executed on the date-based test-set, with skier-specific trackers trained on the corresponding training set.

Performance Measures. To quantify the trackers' performance, we utilized standard measures for long-term tracking evaluations [54]: Precision (Pr ↑), Recall (Re ↑), and F-Score (F-Score ↑). In simple terms, Pr ↑ measures the average number of confidently tracked ground-truth boxes, considering different Intersection-over-Union (IoU) thresholds to determine correct predictions. Re ↑ measures the same number, regardless of the tracker's confidence. F-Score ↑ combines Pr ↑ and Re ↑ into a single aggregated score. We use the Generalized Robustness Score (GSR ↑) [23, 24] to quantify the extent of consecutive frames successfully tracked before losing the target. Such an event occurs whenever the tracker does not resume correct tracking after some time. Additionally, we evaluate the efficiency of trackers by Delay ↓ which measures the time in seconds that has to be waited to obtain the skier's localization after the athlete's performance state is observed in each frame.

6.2. Tracking Impact

Skier localization plays a crucial role in high-level skiing performance analysis pipelines [1, 52, 78, 80]. To assess the impact of trackers in this context, we considered the 2D body and equipment pose estimation task [1, 53]. We focused on the AL and JP disciplines utilizing the Ski2DPose dataset [1] and the YouTube Skijump dataset [53] respectively. These datasets provide sparse 2D key-point annotations for specific frames within a skiing video. We extracted tracking sequences referring to the same skier based on these annotations, and used a tracker to locate the skier in all the frames between the first and last occurrence of the annotations. The predicted boxes from the tracker were then used by a fine-tuned AlphaPose instance [32] to es-

Table 4. **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 →).

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

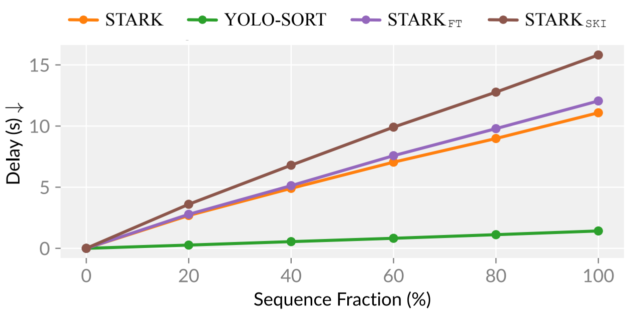

Figure 5. **Waiting time to obtain skier localizations.** The plot illustrates at various fractions of an MC sequence the average time that has to be waited to get the bounding-boxes from the trackers. YOLO-SORT demonstrates the highest efficiency, with minimal delay compared to the actual happening of the skiing performance.

timate 2D poses. We compared the predicted poses with the ground-truth annotations using the Percentage of Correct Key-points (PCK ↑) and the Mean Per Joint Position Error (MPJPE ↓) [1]. We also employed the metrics discussed in Section 6.1 to compare the tracker's predictions with the boxes extracted from the ground-truth key-point coordinates. This analysis allowed us to correlate the tracking and pose estimation accuracies.

7. Results

General Remarks. Table 3 presents the tracking performance of the selected trackers on the date-based test set of SkiTB. Among generic object trackers, STARK results the best. Its performance is comparable to the one achieved on traditional long-term benchmark datasets [30, 48]. These results indicate that generic object trackers struggle to generalize to the application settings represented by SkiTB. on the other hand, skier-specific trackers perform significantly better, with STARK_{FT} improving STARK's F-Score ↑ by 40%. In this score, $\text{STARK}_{\text{SKI}}$ achieves an additional 2% increase over STARK_{FT}. Comparing these findings with the results obtained for the SC setting (available in Table 9 of Appendix D), we observe that camera shotcuts introduce challenges that adversely affect tracking performance. Figure 3 visually illustrates the tracking accuracy of the top methods STARK, YOLO-SORT, STARK_{FT}, $\text{STARK}_{\text{SKI}}$. 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.

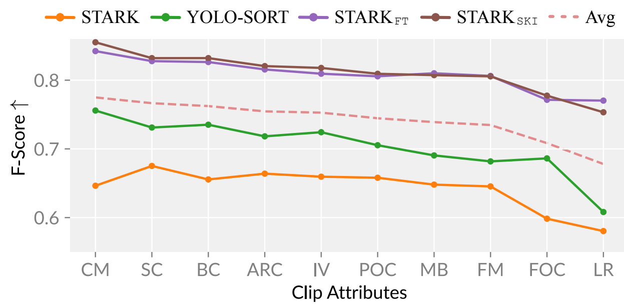

Figure 6. **Tracking performance based on visual attributes.** This plot reports the F-Score ↑ for the different attributes used to characterize the single-camera (SC) clips. We observe that the low resolution (LR), the full occlusion (FOC), and the fast motion (FM) of skiers are the most difficult situations to address.

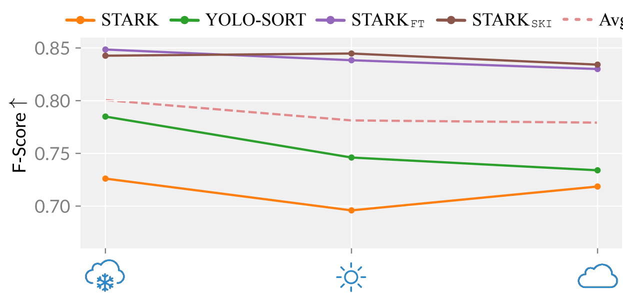

Figure 7. **Impact of the weather conditions on tracking.** This plot displays the F-Score ↑ based on the weather conditions (harsh, sunny, cloudy) characterizing the SC clips. In general, the tracking accuracy is not influenced by different weather conditions.

Table 5. **Tracking after different training conditions.** The table reports STARK_{FT}'s F-Score ↑ under different generalization conditions. 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

In-depth Analysis. In this section, we present a deeper analysis of the application domain on the four most accurate methods presented in Table 3.

JP receives the best tracking performance among the skiing disciplines, followed by AL, while FS presents more challenges. This may be attributed to the complex poses athletes perform in such a discipline, and to the presence of other skiers in the videos as occurs in the sub-disciplines of ski cross and dual moguls.

Figure 4 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. Overall, $\text{STARK}_{\text{SKI}}$ demonstrates the most promising performance. With a recovery time of 1 second or longer, fractions scores exceed 80%. However, shorter recovery time thresholds result in shorter coverages. With a threshold of 1 frame (i.e., 0.03s), $\text{STARK}_{\text{SKI}}$ achieves successful continuous tracking

Table 6. **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. 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	$\text{STARK}_{\texttt{FT}}$	$\text{STARK}_{\texttt{SKI}}$	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

for around the first 40% of the athlete's performance.

Table 4 presents the impact, as reflected in the change in F-Score ↑, of using the YOLOX detector [35, 44] to initialize the tracker. It is observed that the generic STARK is more sensitive to initialization noise. Conversely, skier-specific methods exhibit greater robustness to such initialization, with $\text{STARK}_{\texttt{SKI}}$ maintaining good scores in general. Notably, the detection initialization has a more significant effect on the FS discipline, likely due to the initialization to a wrong skier in multi-athlete sub-disciplines.

In terms of running speed, Figure 5 analyzes the time taken by trackers to provide skier localization at different fractions of the observed skiing performance. YOLO-SORT gives the best efficiency, offering minimal delay compared to the unfolding of the skiing performance. $\text{STARK}_{\texttt{SKI}}$ 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 seconds after his/her performance has ended.

By analyzing the F-Score ↑ per the visual attributes characterizing the SC clips, as reported by Figure 6, we notice that the small size (LR), the full occlusion (FOC), and the fast motion (FM) of skiers are the conditions that determine a performance drop. On the other hand, the camera motion (CM), the scale change (SC), and the background clutter (BC) consist in situations better addressed by the trackers.

Figure 7 shows that trackers work generally at the same accuracy level under different weather conditions. These results demonstrate that such methodologies can be used reliably under challenging image conditions.

Table 5 presents the performance of $\text{STARK}_{\texttt{FT}}$ trained on different splits representing various real-world usage conditions. The results indicate that generalizing to unseen athletes is the most challenging application case. Generalizing to skiing performances occurring after the training ones also proves to be a demanding application condition. Unseen courses instead pose fewer difficulties for generalization.

Finally, Table 6 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 improves the skier tracking results as well as the pose estimation results. For AL on Ski2DPose, $\text{STARK}_{\texttt{SKI}}$ and $\text{STARK}_{\texttt{FT}}$ have nearly the same impact on AlphaPose, while $\text{STARK}_{\texttt{SKI}}$ and YOLO-SORT have comparable impact in JP as represented by the YouTube Skijump dataset. Across disciplines, $\text{STARK}_{\texttt{SKI}}$ results 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. 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 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, and background clutter, and they generalize well to new locations of application. However, the study also identified certain limitations the prevent the methods to be perfect. Challenges arise in maintaining a continuous per-frame reference to the target skier across camera shot-cuts, 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. Future work will focus on addressing these limitations. Solutions may involve refining skier-specific tracking methods, improving generalization, and developing strategies to better integrate with high-level skiing performance understanding modules.

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

References

[1] Roman Bachmann, Jörg Spörri, Pascal Fua, and Helge Rhodin. Motion capture from pan-tilt cameras with unknown orientation. In *International Conference on 3D Vision (3DV)*, 2019. 1, 2, 3, 4, 6, 7, 12, 16

[2] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *CVPR*, 2017. 1

[3] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H.S. Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, 2016. 2

[4] Vinay Bettadapura, Caroline Pantofaru, and Irfan Essa. Leveraging contextual cues for generating basketball highlights. In *ACM MM*, 2016. 1

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 5, 13

[6] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning Discriminative Model Prediction for Tracking. In *ICCV*, 2019. 2

[7] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010. 2, 5

[8] Vasyl Borsuk, Roman Vei, Orest Kupyn, Tetiana Martyniuk, Igor Krashenyi, and Jiři Matas. Fear: Fast, efficient, accurate and robust visual tracker. In *ECCV*, 2022. 5

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[10] Luka Čehovin, Matej Kristan, and Aleš Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *IEEE TPAMI*, 2013. 2

[11] Naga VS Raviteja Chappa, Pha Nguyen, Alexander H. Nelson, Han-Seok Seo, Xin Li, Page Daniel Dobbs, and Khoa Luu. Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition. In *CVPRW*, 2023. 1

[12] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, 2023. 5

[13] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *CVPRW*, 2022. 1

[14] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, 2000. 2

[15] International Olympic Committee. History of Alpine Skiing. 1

[16] The Nielsen Company. FIS Alpine Skiing World Cup Report 2021-22, 2022. 1

[17] The Nielsen Company. FIS Freestyle Ski World Cup Report 2021-22, 2022. 1

[18] The Nielsen Company. Viessmann FIS Ski Jumping World Cup Men 2021-2022, 2022. 1

[19] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, 2022. 2, 5

[20] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-Performance Long-Term Tracking With Meta-Updater. In *CVPR*, 2020. 5

[21] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate Tracking by Overlap Maximization. In *CVPR*, 2019. 2

[22] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *IJCV*, 2021. 13

[23] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Is first person vision challenging for object tracking? In *ICCVW*, 2021. 6, 15

[24] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. *IJCV*, 2023. 2, 5, 6, 12, 13, 15

[25] Matteo Dunnhofer, Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Visual tracking by means of deep reinforcement learning and an expert demonstrator. In *ICCVW*, 2019. 2

[26] Matteo Dunnhofer, Niki Martinel, and Christian Micheloni. Weakly-supervised domain adaptation of deep regression trackers via reinforced knowledge distillation. *IEEE RA-L*, 2021. 2

[27] Matteo Dunnhofer and Christian Micheloni. Cocolot: Combining complementary trackers in long-term visual tracking. In *ICPR*, 2022. 2, 5

[28] Matteo Dunnhofer, Kristian Simonato, and Christian Micheloni. Combining complementary trackers for enhanced long-term visual object tracking. *Image and Vision Computing*, 2022. 5

[29] Matteo Dunnhofer, Luca Sordi, and Christian Micheloni. Visualizing skiers' trajectories in monocular videos. In *CVPRW*, 2023. 15

[30] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking. In *CVPR*, 2019. 2, 5, 7, 12, 13

[31] Heng Fan, Halady Akhilesha Miththanthaya, Harshit, Siranjiv Ramana Rajan, Xiaoqiong Liu, Zhilin Zou, Yuewei Lin, and Haibin Ling. Transparent object tracking benchmark. In *ICCV*, 2021. 2

[32] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE TPAMI*, 2022. 6, 16

[33] Chris Andrew Gadde and C.V. Jawahar. Transductive weakly-supervised player detection using soccer broadcast videos. In *WACV*, 2022. 1

[34] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for Speed: A Benchmark for Higher Frame Rate Object Tracking. In *ICCV*, 2017. 2, 12

[35] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 5, 8, 13, 15

[36] Sam Hare, Stuart Golodetz, Amir Saffari, Vibhav Vineet, Ming Ming Cheng, Stephen L Hicks, and Philip H.S. Torr. Struck: Structured Output Tracking with Kernels. *IEEE TPAMI*, 2016. 2

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[38] David Held, Sebastian Thrun, and Silvio Savarese. Learning to Track at 100 FPS with Deep Regression Networks. In *ECCV*, 2016. 2

[39] Joao F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE TPAMI*, 2015. 2, 5

[40] Yutaro Honda, Rei Kawakami, Ryota Yoshihashi, Kenta Kato, and Takeshi Naemura. Pass receiver prediction in soccer using video and players' trajectories. In *CVPRW*, 2022. 1

[41] S. Hu, X. Zhao, L. Huang, and K. Huang. Global instance tracking: Locating target more like humans. *IEEE TPAMI*, 2023. 3

[42] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE TPAMI*, 2019. 2, 12

[43] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GlobalTrack: A Simple and Strong Baseline for Long-term Tracking. In *AAAI*, 2020. 2, 5

[44] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, Oct. 2020. 5, 6, 8, 15

[45] Maria Koshkina, Hemanth Pidaparthy, and James H. Elder. Contrastive learning for sports video: Unsupervised player classification. In *CVPRW*, 2021. 1

[46] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, et al. The tenth visual object tracking vot2022 challenge results. In *ECCVW*, 2023. 2, 3, 5

[47] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Linbo He, Yushan Zhang, Song Yan, Jinyu Yang, Gustavo Fernández, et al. The eighth visual object tracking vot2020 challenge results. In Adrien Bartoli and Andrea Fusiello, editors, *ECCVW*, 2020. 2, 6, 12, 15

[48] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, Ondrej Drbohlav, Jani Käpylä, Gustav Häger, Song Yan, Jinyu Yang, Zhongqun Zhang, and Gustavo Fernández. The ninth visual object tracking vot2021 challenge results. In *ICCVW*, 2021. 2, 7, 12

[49] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SIAMRPN++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 2, 5

[50] Xin Li and Mooi Choo Chuah. Rehar: Robust and efficient human activity recognition. In *WACV*, 2018. 1

[51] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 1

[52] Katja Ludwig, Philipp Harzig, and Rainer Lienhart. Detecting arbitrary intermediate keypoints for human pose estimation with vision transformers. In *WACVW*, 2022. 1, 2, 6, 12, 15

[53] Katja Ludwig, Daniel Kienzle, Julian Lorenz, and Rainer Lienhart. Detecting arbitrary keypoints on limbs and skis with sparse partly correct segmentation masks. In *WACVW*, 2023. 2, 3, 4, 6, 12, 15, 16

[54] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Performance evaluation methodology for long-term single-object tracking. *IEEE Transactions on Cybernetics*, 2020. 3, 5, 6, 15, 17

[55] Mario Edoardo Maresca and Alfredo Petrosino. MATRIOSKA: A multi-level approach to fast tracking by learning. In *International Conference on Image Analysis and Processing (ICIAP)*, 2013. 2

[56] Seiji Matsumura, Dan Mikami, Naoki Saijo, and Makio Kashino. Spatiotemporal motion synchronization for snowboard big air. In *1st Workshop on Computer Vision for Winter Sports at WACV 2022*. arXiv, 2021. 2

[57] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, 2022. 2

[58] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *ICCV*, 2021. 5

[59] Matthias Mueller, Neil Smith, and Bernard Ghanem. A Benchmark and Simulator for UAV Tracking. In *ECCV*, 2016. 2, 5, 12

[60] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *ECCV*. Springer Verlag, 2018. 2

[61] Hyeonseob Nam and Bohyung Han. Learning Multi-domain Convolutional Neural Networks for Visual Tracking. In *CVPR*, 2016. 2

[62] Hemanth Pidaparthy, Michael H. Dowling, and James H. Elder. Automatic play segmentation of hockey videos. In *CVPRW*, 2021. 1

[63] Hemanth Pidaparthy and James Elder. Keep your eye on the puck: Automatic hockey videography. In *WACV*, 2019. 1

[64] Jiashuo Qi, Dongguang Li, Cong Zhang, and Yu Wang. Alpine skiing tracking method based on deep learning and correlation filter. *IEEE Access*, 2022. 2

[65] Julian Quiroga, Henry Carrillo, Edisson Maldonado, John Ruiz, and Luis M. Zapata. As seen on tv: Automatic basketball video production using gaussian-based actionness and game states recognition. In *CVPRW*, 2020. 1

[66] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1, 2, 3, 6, 16

[67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 1, 2, 3, 16

[68] Helge Rhodin, Frédéric Meyer, Jörg Spörri, Erich Müller, Victor Constantin, Pascal Fua, Isinsu Katircioglu, and Mathieu Salzmann. Learning monocular 3d human pose estimation from multi-view images. In *CVPR*, 2018. 3

[69] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020. 5

[70] International Ski and Snowboard Federation. International Ski and Snowboard Federation. 1, 4, 5, 13

[71] Philip Steinkellner and Klaus Schöffmann. Evaluation of object detection systems and video tracking in skiing videos. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2021. 2, 3, 4

[72] Jonas Theiner and Ralph Ewerth. Tvcalib: Camera calibration for sports field registration in soccer. In *WACV*, 2023. 1

[73] Jonas Theiner, Wolfgang Gritz, Eric Müller-Budack, Robert Rein, Daniel Memmert, and Ralph Ewerth. Extraction of positional player data from broadcast soccer videos. In *WACV*, 2022. 1

[74] Laurent Vanat. 2022 International Report on Snow & Mountain Tourism, 2022. 1

[75] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *CVPRW*, 2022. 1

[76] Kanav Vats, Mehrnaz Fani, David A. Clausi, and John Zelek. Puck localization and multi-task event recognition in broadcast hockey videos. In *CVPRW*, 2021. 1

[77] Kanav Vats, William McNally, Pascale Walters, David A. Clausi, and John S. Zelek. Ice hockey player identification via transformers and weakly supervised learning. In *CVPRW*, 2022. 1

[78] Dejan Štepec and Danijel Skočaj. Video-based ski jump style scoring from pose trajectory. In *WACVW*, 2022. 1, 2, 4, 6, 12, 15

[79] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *CVPR*, 2021. 16

[80] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *ACM MM*, 2019. 1, 2, 6, 15

[81] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE TPAMI*, 2015. 2, 5, 12, 13, 15, 17

[82] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021. 2, 5, 14

[83] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'Skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In *ICCV*, 2019. 2, 5

[84] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 2, 5

[85] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *CVPR*, 2017. 2

[86] Yulin Zhu and Wei Qi Yan. Ski fall detection from digital images using deep learning. In *International Conference on Control and Computer Vision*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[87] Michael Zwölfer, Dieter Heinrich, Kurt Schindelwig, Bastian Wandt, Helge Rhodin, Joerg Spoerri, and Werner Nachbauer. 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*. arXiv, 2021. 2