IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

AI-Based Cropping of Ice Hockey Videos for Different Social Media Representations

MEHDI HOUSHMAND SARKHOOSH[1,3,*], SAYED MOHAMMAD MAJIDI DORCHEH[1,3,*], CISE MIDOGLU[2,3], SAEED SHAFIEE SABET[3], TOMAS KUPKA[3], DAG JOHANSEN[4], MICHAEL A. RIEGLER[1,2], and PÅL HALVORSEN[1,2,3]

[1]Oslo Metropolitan University, Oslo, Norway
[2]SimulaMet, Oslo, Norway
[3]Forzasys AS, Oslo, Norway
[4]UiT The Arctic University of Norway, Tromsø, Norway

Corresponding author: Mehdi Houshmand Sarkhoosh (e-mail: mehdi@forzasys.com).

*Both authors contributed equally to this research.

ABSTRACT Sports multimedia is among the most prominent types of content distributed across social media today, and the retargeting of videos for diverse aspect ratios is essential for a suitable representation on different social media platforms. In this respect, ice hockey is quite challenging due to its agile movement pattern and speed, and because the main reference point (puck) is very small. In this paper, we introduce a novel pipeline for intelligent video cropping tailored for ice hockey. Our main goal is to identify regions of interest in video frames by detecting and tracking the hockey puck using state-of-the-art AI models. Our pipeline employs scene detection, object detection, outlier detection, and smoothing as key features. Our proposed pipeline called SmartCrop-H is not only highly efficient and configurable with respect to target aspect ratios, but also addresses the automation needs in this domain. Our comprehensive evaluation, comprising objective and subjective measures, shows the overall efficiency of the entire pipeline, including assessments of both the individual components and the end-to-end system performance. We also demonstrate the practical applicability of SmartCrop-H with a user study, which indicates that our framework performs on par with, or even surpasses, professional tools in terms of output quality.

INDEX TERMS AI, aspect ratio, cropping, ice hockey, social media, video processing

I. INTRODUCTION

IN today's fast-paced digital landscape, the consumption of media content is no longer confined to traditional platforms. Ice hockey, with its intense fan base, epitomizes the type of content fans eagerly consume on a variety of devices, ranging from expansive television screens to handheld smartphones. Each viewing platform, with its unique aspect ratio, demands a customized presentation to ensure consistent content delivery to audiences, regardless of their viewing device [1].

Curation of multimedia content for various target devices and platforms is traditionally a tedious manual job, despite and in conflict with the expectation of fast publishing [2]. Conventionally, video cropping has been undertaken using tools, such as Adobe Premiere Pro [3] and Final Cut Pro [4], which require laborious frame-by-frame editing. However, this manual approach does not meet the needs for real-time broadcasting and voluminous content. The industry has thus pivoted towards automated solutions, with emerging deep learning models promising more efficient video retargeting. However, the dynamic and unpredictable nature of ice hockey broadcasts brings forth unique challenges: tracking and maintaining the visibility of the smaller and faster focal object (puck) across varying aspect ratios is more demanding than in other sports domains, where this is already a demonstrably hard problem [5].

In this paper, we aim to automate the curation of ice hockey highlights for direct publication on social media. We introduce SmartCrop-H, an end-to-end pipeline optimized for cropping ice hockey videos into various aspect ratios tailored for different social media platforms. Based on the idea proposed for cropping soccer videos [6], our solution is designed to ensure that the hockey puck remains in clear view, thus enhancing the viewing experience for fans worldwide. However, as the previous soccer approach does not work on ice hockey videos due to domain-specific challenges (e.g., small puck, faster pace, different broadcast settings, etc.), SmartCrop-H reworks the fundamental ideas for scene detection, object

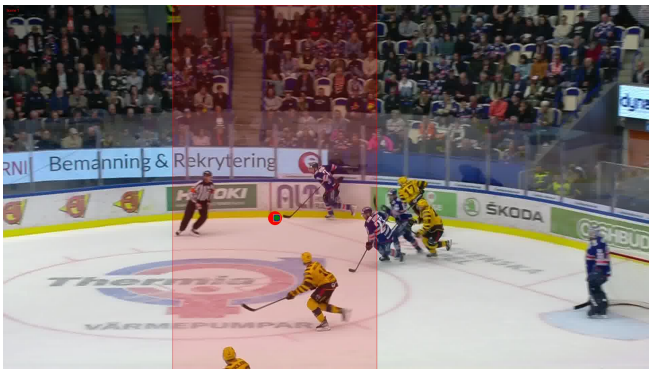


FIGURE 1: Cropping from 16:9 to 9:16 aspect ratio using object detection. Red dot = cropping-center or POI, red transparent square = cropping area or ROI.

detection, and outlier detection, along with adding new logic for smoothing, to calculate the optimal cropping center for video frames (Figure 1).

The contributions of this paper are as follows:

- An implementation of an end-to-end automated video cropping pipeline customized for ice hockey, which possesses distinct properties compared to other sports such as soccer.
- A comprehensive objective evaluation of the pipeline, assessing both individual component performance and overall system performance.
- A subjective evaluation of the end-to-end pipeline, where user study participants demonstrated a preference for SmartCrop-H over alternative cropping methods.
- A detailed competitor analysis conducted through a separate user study, comparing SmartCrop-H with Final Cut Pro Auto Reframe, Adobe Premiere Pro Auto Reframe, manual cropping using Adobe Premiere Pro, and static cropping.

The results from objective and subjective experiments, with a total of 64 participants in 3 different user studies, suggest that SmartCrop-H improves end-users' Quality of Experience (QoE) and is suitable for real-world deployment. Compared to other tools, SmartCrop-H shows significant superiority over basic and less sophisticated methods, marginal superiority over automated techniques, and highlights areas for further development to reach the standards of manual cropping.

II. BACKGROUND

A. VIDEO ASPECT RATIO ADJUSTMENT

Algorithms for adjusting the video aspect ratio are essential in maintaining content quality on different viewing devices. Content-adaptive reshaping (warping) focuses on selectively altering certain areas of an image while preserving key regions using a grid-based scaling approach [7], [8]. Segment-based exclusion (cropping) targets a particular image or frame area, ignoring elements outside the selected boundary [9]–[12]. Seam extraction identifies and removes non-essential pixel lines, optimizing the image for different aspect ratios [13], [14]. Hybrid methods combine these foundational strategies, such as merging cropping with content-adaptive reshaping [15], or integrating seam extraction with segment-based exclusion [16]. A landmark method by Apostolidis and Mezaris [17] uses cropping to retarget videos to different aspect ratios, with a focus on minimizing semantic distortions. These methods, while adept at general video aspect ratio transformation, may not be ideal for scenarios like sports videos where tracking a specific object like a ball or puck is crucial. This limitation arises because the algorithm focuses on areas of high visual saliency, which may not always include the target object. Consequently, this approach might exclude the object of interest during cropping. Additionally, the method prioritizes minimizing semantic distortions, which differs from the needs of videos where object presence is more critical than overall image quality or semantic content.

By enabling flexible and intelligent adaptation of video content to various screen sizes, these algorithms play a pivotal role in bridging the gap between diverse digital platforms and viewing devices, ensuring that every viewer enjoys an optimal visual experience, a necessity in our increasingly screen-oriented world.

B. PUCK DETECTION AND TRACKING IN ICE HOCKEY

Advancements in computer vision and computing technologies have significantly impacted the field of sports analytics, including the localization and tracking of the puck in ice hockey. Various methods, including deep network regressors and hierarchical graph-based methods, have been developed to address the challenges posed by the puck's small size, rapid movement, and occlusions in broadcast ice hockey games [18]–[23]. These advances have not only enhanced the accuracy of puck tracking systems, but have also contributed to a deeper understanding of the game's dynamics.

A notable contribution in this domain is the implementation of an automatic system for puck tracking and play localization [20]. Utilizing the puck as a key indicator of action, this system employs advanced techniques such as deep network regressors trained on high-definition video. The authors demonstrate the effectiveness of their method by providing a cost-effective solution for dynamic video capture and retargeting, particularly in amateur-level hockey. The impact of puck tracking technology extends to professional sports broadcasting as well, as evidenced by its use in Fox Sports' NHL coverage. Despite facing initial challenges, technology has led to significant advancements, including patents and improved viewer ratings and engagement [24]. An advanced system has been developed for determining puck possession and location, which incorporates two innovative modules: one for direct puck candidate detection and another for estimating puck location based on player motion fields. This system represents a major step forward in automated hockey video analysis, providing detailed strategies for puck detection and possession [25]. Research using indoor localization data from the Wisehockey platform has focused on detecting bodychecks in ice hockey. Employing a random

forest algorithm, this study achieved high accuracy but also highlighted the need for further improvements to reduce false positives. Future enhancements could include the integration of acceleration data and video analysis for more accurate detection [26].

While existing methodologies are robust, they often require high-definition video input and are tailored for professional broadcast environments, which may not be available or practical for all applications. The complexity and specificity of these methods limit their generalizability and scalability, particularly for amateur settings or when computational resources are constrained. In the context of hockey broadcasts, the application of object detection models presents unique challenges, especially concerning the consistent visibility of the puck, a central element in the sport. Among the primary difficulties is **inconsistent detection**. Factors such as swift changes in lighting, obstruction by players, and varying camera angles can significantly affect the accurate detection of the puck, despite using sophisticated models. Moreover, **adapting to different aspect ratios** while keeping the puck visible, undistorted, and centrally located is not straightforward. Techniques like content-adaptive reshaping or seam extraction can sometimes reduce visibility if not executed correctly. Additionally, implementing outlier detection and smooth-function techniques to **rectify detection errors** adds to the complexity. While these methods aim to increase precision, they may paradoxically introduce inaccuracies due to the fast-paced and unpredictable nature of hockey.

The You Only Look Once (YOLO) framework, a revolutionary approach to object detection, streamlines the process by performing detection in a single pass through the neural network, thus significantly speeding up the process [27], [28]. Its unique design allows for real-time detection, making it highly efficient and suitable for applications where speed is crucial. Taking advantage of these strengths, we have integrated YOLO into our SmartCrop-H pipeline. YOLO's adaptability to different environments and its capability to handle scenarios like rapid movement and small object sizes, such as tracking a hockey puck, are particularly beneficial. Furthermore, its widespread use in various domains highlights a strong support community and ongoing development, which are essential for the continuous improvement of our puck-tracking system. YOLO's balance of high performance, ease of implementation, and adaptability align with the diverse requirements of our project, as detailed in Section III.

C. COMMERCIAL APPLICATIONS

In the field of AI-driven sports media solutions, companies such as Magnifi [29], Pendular [30], Backlight [31], and WSC Sports [32] are prominent for their contributions in personalized highlight generation and video aspect ratio adjustments, which are key to enhancing viewer engagement on various platforms. Despite their advancements, these solutions often lack aspects critical for real-life deployments, such as lightweight operation and live edge processing, essential in dynamic sports settings. Moreover, popular video editing software tools such as Adobe Premiere Pro and Final Cut Pro, despite being feature-rich, are not primarily tailored for sports applications, facing challenges in handling fast-paced sequences and intricate color contrasts seen in sports like ice hockey. This points to a noticeable gap in the market for specialized technologies that cater specifically to the unique dynamics and requirements of ice hockey, particularly in capturing and emphasizing pivotal moments in this swift-paced sport.

D. SMARTCROP PIPELINE

In previous work [6], we proposed a cropping pipeline tailored for soccer called SmartCrop, which relies on the detection of the soccer ball for the determination of the point of interest (POI). However, we faced challenges when we ran this pipeline on video clips from ice hockey game broadcasts. In this paper, we propose a modified pipeline called SmartCrop-H specifically designed to address the different properties of ice hockey, including:

- **Game pace:** Ice hockey has a faster pace than soccer leading to quicker camera movements.
- **Shot (scene) types:** Ice hockey broadcasts have more close-up and medium shots compared to soccer, which has more long and full shots.
- **Color contrasts:** The detection of the hockey puck against the white ice background is relatively more challenging, especially considering the game's swift pace and the presence of other black markings and lines on the ice. This contrasts with the visibility of a soccer ball against the green pitch, which is generally more apparent when the ball is on the ground. However, this visibility can vary when the soccer ball is held by a player during a throw-in or by the goalkeeper, or when it is in the air with potentially confusing backgrounds like the players' heads.
- **Regions of interest:** Although important events tend to be around the hockey puck in general, they might also appear in the absence of the puck (e.g., player huddle), as opposed to soccer where the main action is almost always centered around the soccer ball (with the partial exception of offsides).
- **Broadcast video properties:** Frame rate for soccer broadcasts is around 25 fps, whereas it is around 50fps for hockey; resolution tends to be higher for hockey (e.g., 1920 for our particular content source, as opposed to the 1280 mentioned in [6]).

Compared to the original SmartCrop pipeline for soccer, SmartCrop-H updates or replaces most of the pipeline modules to meet the unique properties of ice hockey. For the faster pace of ice hockey, we have introduced a smoothing module. This module smoothly adjusts the POI, preventing abrupt changes in the cropping window that could disturb the viewer, while still capturing the essence of the frame. Regarding shot (scene) types, we have implemented an outlier detection module. This module recognizes the distribution of

objects in medium shots, where elements can be more widely spread in comparison to long shots, where important objects such as the ball or puck are typically not located at the edges of the frame. To address the color contrast challenge, we have fine-tuned a YOLO model specifically for puck detection. We have optimized the scene detection module to better differentiate between the scenarios in ice hockey and soccer, taking into account the distinct visual elements of each sport. Additionally, we present a comprehensive system performance analysis with a focus on optimized resource usage, so that our pipeline can support the more demanding broadcast video properties of ice hockey, which was not available in previous work. These targeted enhancements ensure that SmartCrop-H is precisely tailored to the dynamic and visual intricacies of ice hockey broadcasting.

III. PROPOSED FRAMEWORK

The fundamental principle driving the SmartCrop-H pipeline, similar to the original SmartCrop pipeline, is for the POI to be used as the center point of the cropping area (Figure 1). In the ice hockey scenario, the puck serves as the POI. The selection of the puck as the POI is based on several key considerations and methods. Firstly, extensive video analysis of ice hockey games was conducted by authors to understand the dynamics and movement patterns of the game. This analysis involved tracking the puck's position and evaluating its relevance to the overall gameplay and viewer focus. Additionally, input from ice hockey experts, as well as literature reviews, were sought to confirm the puck's critical role as the central element in the sport [19]. Moreover, the determination of POI extends beyond hockey to various sports contexts. In volleyball [33], for instance, the ball is typically the POI due to its continuous movement and centrality in play actions. Similarly, in soccer, the ball also serves as the primary POI [6], although in some cases, key players or specific regions of the field could be considered based on the tactical analysis of the game [34]. These determinations involve a combination of video analysis, expert consultation, and algorithmic tracking to ensure that the most relevant and action-centric POI is identified for effective cropping. Hence, when the puck is visible within a frame, we use it as the primary focal point for the cropping. When it is absent, we seek an appropriate alternative focal point by employing smoothed interpolation or relying on frame-centered cropping as a fallback.

As depicted in Figure 2, the SmartCrop-H pipeline consists of 7 modules with intermediate logic in between. These are: (1) Pre-processing module, (2) Scene detection module, (3) Object detection module, (4) Outlier detection module, (5) Smoothing module, (6) Cropping module, (7) Post-processing module. The pipeline takes as input an HLS playlist URL or an `mp4` file and outputs an `mp4` file. A demonstration of the SmartCrop-H pipeline is provided in [35].

A. PRE-PROCESSING MODULE

Our pipeline supports various video formats, with HTTP Live Streaming (HLS) as the default input format. To optimize object detection accuracy, the pipeline analyzes the HLS manifest to select the highest quality stream available. Swedish Hockey League (SHL) [36] broadcasts are offered in two resolution options: 854×480 and 1920×1080. The choice of resolution affects the processing speed and detection accuracy. Higher resolution frames enhance detection precision but may increase processing time, while lower resolutions may lead to false positives or negatives [37]. The stream undergoes conversion to H.264 encoded videos within an `mp4` container before being processed by the object and scene detection modules.

B. SCENE DETECTION MODULE

Understanding the changes in the camera view that occur during broadcasting requires defining scenes, i.e., a continuous video sequence from the same camera view, and thus, each such scene denotes a change in camera angle. By distinguishing between scenes such as wide shots, which capture most players and the puck, and close-up shots, which focus solely on player faces, we establish our cropping logic. Scenes where the puck occupies a threshold percentage of frames prompt puck tracking as the POI, while others default to frame-centered cropping. The latter approach is derived from the implicit convention in sports video recording, where cameramen typically center the most significant object, usually a player's face or group of players, within the center of frames. SmartCrop-H integrates two methods to establish a robust scene detection module: TransNetV2 [38] (a machine learning model) and PySceneDetect [39], [40] (a Python library). TransNetV2 specializes in identifying individual scenes, a feature that our pipeline leverages for context-sensitive video cropping. Initially, we started with TransNetV2, but through our analysis, we identified some misclassifications in identifying scene changes. This issue could be attributed to the specific characteristics of ice hockey video footage, which is significantly different from the datasets like ClipShots [41] and RAI used to train TransNetV2. The unique dynamics, lighting conditions, and rapid movements in hockey videos present challenges that these pre-trained datasets do not encompass. Recognizing the need for an additional level of ca-

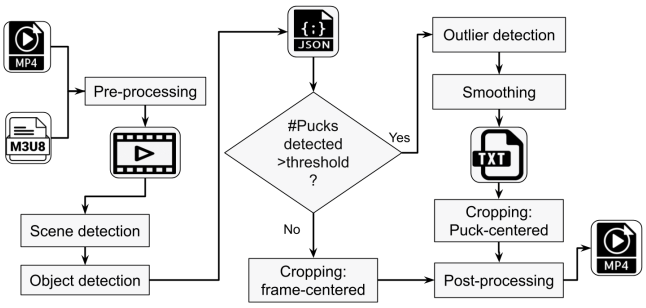


FIGURE 2: Pipeline overview.

	Model	#Par (M)	GFLOPs	FPS	AP (val)
1	SSD [43]	34	100	59	50.3
2	Faster R-CNN [44]	42	180	50	53.2
3	MobileNets with SSD [43]	20	90	65	48.9
4	**YOLO [45]**	**50**	**140**	**75**	**52.5**

TABLE 1: Comparative metrics for real-time object detectors.

	YOLOv8 Ver.	#Par(M)	GFLOPs	FPS	APval	APval50	APval75
1	Nano	5	30	90	40.0	60.0	42.0
2	Small	15	80	75	45.0	65.0	48.0
3	**Medium**	**30**	**140**	**65**	**50.0**	**68.0**	**53.0**
4	Large	43	165	60	52.9	69.8	57.5
5	X-Large	68	257	50	53.9	71.0	58.7

TABLE 2: Comparative metrics for different versions of YOLOv8 [46].

pability in scene identification, we incorporated PySceneDetect. PySceneDetect utilizes histogram-based methods and content changes to detect scenes, which can complement the machine learning approach of TransNetV2. Specifically, PySceneDetect considers changes in brightness, color histograms, and pixel values, making it particularly effective for the high-speed, high-contrast environment of hockey games. The integration of PySceneDetect in SmartCrop-H is crucial for handling these unique challenges, ensuring more accurate scene detection and thereby improving the overall performance of the video cropping pipeline.

C. OBJECT DETECTION MODULE

As described in the previous section, the identification of the puck in the rink will be a crucial factor in developing the cropping logic. The purpose of this module is to detect the ice hockey puck in each frame. If the number of detected pucks surpasses the given threshold compared to the total number of frames in that scene, the puck will serve as the POI.

The selection process for the appropriate object detection model involved evaluating classical methods and machine learning-based methods. Classical approaches such as Haar Cascades [42], while foundational, lack efficiency in complex scenarios. Machine learning methods like SVMs improve accuracy but are less efficient than deep learning methods. Deep learning approaches, especially end-to-end and real-time object detection models such as RT-DETR-R50, provide advanced solutions by leveraging neural networks for automatic feature learning. These End-to-end models offer high accuracy but are less suitable for real-time applications due to their processing time. Real-time models, essential for applications requiring rapid feedback, process video feeds almost instantaneously. YOLO models, in particular, process the entire image in a single evaluation, significantly reducing computation time. In reviewing the metrics for various real-time object detection models (Table 1), YOLO models demonstrated a superior blend of speed and precision, making them ideal for high-speed object detection environments like ice hockey video.

Given the superior blend of speed and precision (Table 1), YOLO was selected for real-time detection. Among the YOLOv8 versions, YOLOv8-Medium emerged as the optimal choice due to its balanced profile: 30 million parameters, 140 GFLOPs, 65 FPS, and high precision (Table 2). YOLOv8-Medium balances performance and efficiency, making it ideal for real-time ice hockey video analysis. It meets the demands of dynamic environments by providing both speed and accuracy.

1) Motivation for Hockey-Specific Model

As shown in Figure 3, when we investigated the performance of existing object detection models, we observed that the general checkpoint of YOLOv8 Medium model and the Y8m_sc11 model (custom YOLOv8 version fine-tuned, for soccer [6]) had limitations in puck detection. These models, optimized for broader sports contexts, showed inadequate sensitivity to the unique characteristics of the ice hockey puck. In the first column of Figure 3, YOLOv8 Medium is expected to detect all objects and Y8m_sc11 to detect three classes (ball, player, logo). The red bounding boxes in the first and second columns indicate objects that the models failed to detect.

2) Dataset Curation and Model Training

In our pursuit to improve the SmartCrop-H pipeline, particularly for accurate puck tracking, we used a dataset of 800 annotated images from the Swedish Hockey League (SHL) [36]. Leveraging this dataset, we developed the custom Y8_sc_m model, tailored to meet the specific requirements of our application. This model underwent a training process spanning 100 epochs. It features Non-Maximum Suppression (NMS) thresholding, optimized for image processing at a resolution of 1280 pixels. The training protocol involved a batch size of 16 and a patience setting of 100 epochs, utilizing the AdamW optimizer to ensure efficient learning. The model's learning rate was initially set at 0.001, gradually increasing to a final value of 0.02, with a dropout rate of 0.5 applied to prevent overfitting, thereby ensuring a robust and effective training outcome. The model weights are publicly available on GitHub [47].

We utilized the Labelbox [48] platform to annotate our dataset. Our annotation strategy was focused solely on a single class, i.e., the puck. The dataset encompasses a wide range of game situations to ensure robust puck detection, including:

- **Fast breaks and counter-attacks:** Capturing frames where the puck is rapidly transitioned from defense to offense, challenging the model to detect the puck in high-speed scenarios.
- **Close-up scrimmages:** Including images where players are in close proximity around the puck, often leading to

(a) Sample 1: YOLOv8m (b) Sample 1: Y8_sc11 (c) Sample 1: Y8_sc-m

(d) Sample 2: YOLOv8m (e) Sample 2: Y8_sc11 (f) Sample 2: Y8_sc-m

(g) Sample 3: YOLOv8m (h) Sample 3: Y8_sc11 (i) Sample 3: Y8_sc-m

FIGURE 3: Object detection performance (visual comparison): YOLOv8-Medium vs. custom models for soccer and hockey.

obscured views of the puck.

- **Long shots and wide Angles:** Frames that show the gameplay and the puck from a distance, testing the model's ability to detect smaller objects.
- **Goal scoring moments:** High-intensity frames capturing the moments of scoring, where the puck is often in the net or close to goalkeepers.
- **Power plays and penalty kills:** Situations where team formations are different due to penalties, affecting the puck's visibility and trajectory.
- **Face-offs:** Including the start of play scenarios where the puck is dropped between two players, providing varied orientations and positions.
- **Board plays:** Capturing the puck in play along the boards, where it can be partially obscured or in tight spaces.
- **Puck in flight:** Frames where the puck is airborne, presenting different angles and shadows compared to when it is on the ice.

The third column of Figure 3 illustrates the Y8_sc_m model's remarkable proficiency in puck detection. This highlights the importance of specialized model training for sports applications. While general-purpose models have broad utility, fine-tuning with domain-specific data, as in the example of the Y8_sc_m model, significantly enhances effectiveness.

D. OUTLIER DETECTION MODULE

Recognizing that the object detection from the previous module, which predicts puck positions in each frame, may contain outliers, this module employs various techniques to identify and remove anomalous data points. Outliers refer to cases where the object detection falsely identifies a puck when none is present, known as a false positive. Additionally, outliers occur when the object detection correctly identifies multiple pucks, but only one is the main puck of interest (see figure 4). For instance, a puck may be correctly detected at the center of the play while others are outside the play area. Moreover, outliers could occur when a puck is detected in areas where it is not typically found, such as the corner areas of each frame. Here, we will discuss methods for addressing these outliers.

1) Outlier Detection Methods

We incorporate three primary outlier detection methods to enhance the robustness of our system.

Z-score method: Outliers are identified by calculating the Z-score using Eq. 1, where z is the Z-score, x is the data point, μ is the mean of the data set, and σ is the standard deviation of the data set. A data point x is considered an outlier if $|z| >$ threshold, commonly set to 2 or 3 [49].

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Modified Z-score method: An outlier is identified based on the median and median absolute deviation (MAD) using

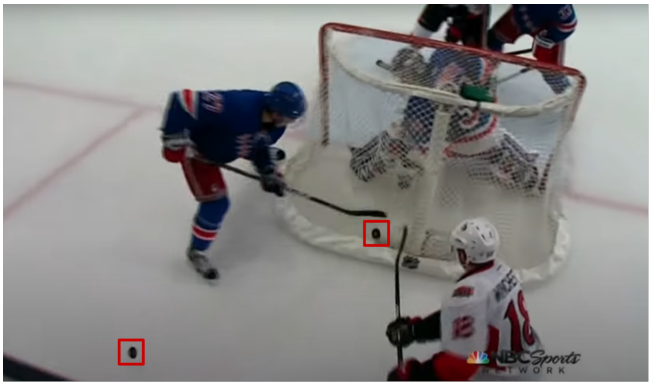


FIGURE 4: An example of a situation where 2 pucks correctly being identified on a ice-hockey rink. Pucks have been shown with a red box around them.

Eq. 2, Eq. 3, and Eq. 4, where x is the data point, M is the median of the data set, MAD is the median absolute deviation, and α is a threshold, often set to 2 or 3. In this method, x is an outlier if its scaled distance from M exceeds $\alpha \times MAD$ [49]. The choice of threshold values (α) such as 2 or 3 is guided by balancing sensitivity and specificity in outlier detection. A threshold of 2 is more sensitive and may identify more outliers, which is useful in datasets where it is critical to capture most anomalies. Conversely, a threshold of 3 is more conservative, reducing the risk of false positives, and is preferable in contexts where outlier misidentification could lead to significant errors. These thresholds are empirically derived from statistical properties of the MAD and validated through simulations and practical applications across various datasets [50].

$$M = \text{Median}(\{x_1, x_2, \ldots, x_n\}) \tag{2}$$

$$\text{MAD} = \text{Median}(\{|x_1 - M|, |x_2 - M|, \ldots, |x_n - M|\}) \tag{3}$$

$$\text{Outlier if } |x - M| > \alpha \cdot \text{MAD} \tag{4}$$

Interquartile Range (IQR) method: Outliers are identified using quartiles. IQR is calculated by Eq. 5, where $Q1$ is the first quartile and $Q3$ is the third quartile. Outliers are defined as values outside the range given by Eq. 6, where k is a scaling factor, often 1.5 for mild outliers or 3 for extreme outliers [49].

$$\text{IQR} = Q3 - Q1 \tag{5}$$

$$\text{Outlier if } [Q1 - k \cdot \text{IQR}, Q3 + k \cdot \text{IQR}] \tag{6}$$

2) Outlier Detection for Soccer vs. Hockey

To enhance our understanding of what constitutes an outlier in hockey videos, we conducted an examination of the puck position distribution patterns across sample video frames. The examination allowed us to make a comparison between the hockey puck and soccer ball behavior. Ground truth on the spatial positions of the hockey puck and soccer ball was obtained from approximately 700 sample video frames for each.

Figure 6 presents the probability distribution of the normalized x-positions and y-positions of pucks and balls, as estimated by a Kernel Density Estimate (KDE) [51]. It shows the likelihood of finding the puck or ball at various points along the x and y axis of the video frame. The density here refers to the estimated probability density function, which provides a smooth curve indicating where data points (puck or ball positions) are concentrated. A high value of the density, such as 3, indicates a high concentration of positions around that normalized x or y position, meaning it is a common location for the object in the data set. It is important to note that while probability density values can exceed 1, they are not probabilities themselves but rather indicate relative concentration.

A key finding, as illustrated in Figure 6, is the disparity in puck and ball positions along the Y-axis. This variance can be attributed to the distinct shooting mechanisms employed in hockey and soccer broadcasts (and the corresponding difference in the dominant scene types, see Figure 5). Notably, in hockey, the puck distribution is concentrated within the initial 30% of the Y-axis range. In contrast, the X-axis distribution of the puck position mirrors that of the soccer ball more closely.

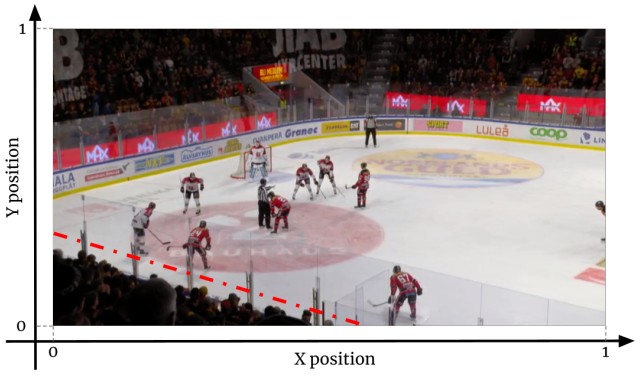
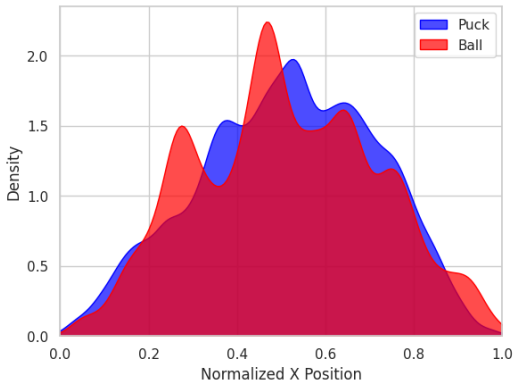
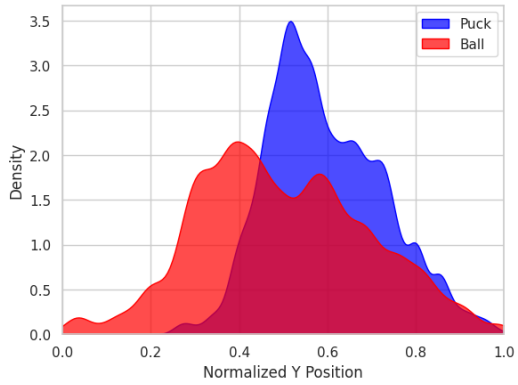


FIGURE 5: Sample frame from an ice hockey broadcast showing fans in the initial 20% of the y-axis, puck detections in this region are excluded in the outlier detection module.

Drawing from our analysis of the puck position distribution in hockey videos, enhancements have been made to the outlier detection algorithm within our pipeline. A notable adaptation is the introduction of a Y-axis threshold, specifically focusing on the top 70% of the video frame. This decision comes from our findings that most puck positions are located predominantly within this upper 70% segment of the Y-axis, as illustrated in Figure 6. This pattern is distinctly different from soccer, where the ball distribution is more uniformly spread along the Y-axis. Visually, this implies that most critical hockey actions, such as key puck movements and interactions,



(a) Normalized X position



(b) Normalized Y position

FIGURE 6: Comparative visual analysis of soccer ball and hockey puck X and Y positions.

occur predominantly in the upper part of the frame. Due to these differences, we need different system configurations for the two different sports.

E. SMOOTHING MODULE

The smoothing module enhances the detection and tracking accuracy of the puck in ice hockey videos by reducing noise and erratic fluctuations in the detected positions. This leads to more stable and reliable tracking of the puck over time, which is essential in the fast-paced and dynamic environment of ice hockey. To streamline the transition between puck positions (POIs) in the video processing pipeline, an Exponential Moving Average (EMA) has been integrated. EMA, a method for smoothing data series, emphasizes recent observations more than older ones. EMA uses *weighted averages with exponentially decreasing weights*, thus assigning more importance to the latest data and less to older observations. The selection of coefficients $(\alpha, 1-\alpha)$ for the EMA calculation allows for a flexible weighting system that prioritizes recent data points while still considering historical data. EMA is computed using Eq. 7, where EMA_t is the EMA value at time t, P_t is the actual data point at time t, EMA_{t-1} is the EMA value at time $t-1$, and α is the smoothing factor which lies in the range $(0, 1)$.

$$\mathrm{EMA}_t = \alpha \cdot P_t + (1-\alpha) \cdot \mathrm{EMA}_{t-1} \tag{7}$$

The parameter α dictates the degree of weighting reduction; a higher α means that more emphasis is placed on the recent data points, rendering the EMA more responsive to new information. In contrast, a smaller α value extends the influence of older data points, producing a more stable EMA over time. In essence, the EMA achieves a balance between the immediacy of the latest data point and the continuity of historical data, culminating in a refined output. To determine the optimal α value, a subjective smoothing study (described in Section IV-F) was conducted. For this smoothing module, α parameter was experimentally selected, ensuring that the chosen α value provides the best balance between responsiveness to new data and the stability of the smoothed series. This approach ensures that the EMA is tailored to the specific needs of the video processing pipeline, enhancing its performance in transitioning between POIs.

F. CROPPING MODULE

Finally, cropping is used to isolate the Region of Interest (ROI) within each frame in videos. This module crops each frame-based ROI, which is calculated by POI from the previous modules and also the *Aspect Ratio* parameter in the pipeline configuration to produce the requested output size (see Figure 1). The calculated POIs are divided into two categories based on each scene in our pipeline:

1) **Frame-centered cropping**: The cropping center is statically selected in the middle of the frame because there are not enough pucks detected to consider tracking and following.
2) **Puck-centered cropping**: The cropping center is based on the coordinates of the puck as detected and calculated by the detection and smoothing modules.

These POIs are coupled with the designated aspect ratio to delineate the ROIs. For instance, for an aspect ratio of 1:1, the POIs necessitate placement in the middle of the region, taking into account equal proportions on each side. In a two-dimensional context, such as cropping an image or video frame, these sides represent the top, bottom, left, and right sides, respectively, to ascertain the specific portion for cropping. This methodology ensures balanced cropping, maintaining visual symmetry and focus on the ROI. This approach varies for other aspect ratios, such as 9:16, where the POIs are situated in the region's middle but adjusted for a horizontal aspect ratio of 9 and a vertical aspect ratio of 16 to determine the ROIs and execute the cropping.

G. POST-PROCESSING MODULE

The cropped frames from the previous section serve as the input to the Post-Processing module. Its primary function is to create an `mp4` file, returning this file as the output of the pipeline in the aspect ratio selected in the preceding module. Additionally, this module can incorporate other visualization

functionalities, such as overlaying logos, graphics or other elements onto the output video.

IV. EVALUATION PER MODULE

The goal of the objective evaluation is to investigate the performance of each module in the SmartCrop-H pipeline, and determine the best models or methods to be used in each module in the final version (to be deployed in production).

A. DATASETS FOR EVALUATION

Scene detection module: For the evaluation of the scene detection module, we annotated five hockey videos. These videos collectively comprise approximately 30,000 frames. Each frame was carefully examined to identify and mark the scene changes with precise ground truth annotations, as illustrated in Figure 7. Our criteria for a scene change encompassed both camera transitions and variations in camera zoom levels. The videos were intentionally selected based on their dynamic nature, featuring frequent scene changes to rigorously test the detection module. In Figure 7, we present the first frames of four initial scenes from one of these videos. These frames were annotated as indicators of scene changes, showcasing the varied scenarios that our module needs to identify. Notably, each frame represents the commencement of a new, consecutive scene (see Section IV-C).

Outlier detection module: For the evaluation of the outlier detection module, we used 700 ice hockey frames, where the puck's position being manually annotated frame by frame as ground truth.

B. EVALUATION METRICS

We use the F1 score to evaluate the performance of the scene detection module, which provides a balanced measure of precision and recall (more specifically, the harmonic mean of precision and recall). To evaluate the performance of the object detection module, we assess the accuracy of alternative object detection models in terms of the count of True Positive (TP) and False Negative (FN) samples. We use Mean Absolute Error (MAE) to evaluate the performance of the outlier detection module. MAE is given by Eq. 8, where n is the number of observations, y_i represents the actual puck position in each frame (ground truth), and $\hat{y}_i$ is the puck position post-outlier detection. This method provides a clear evaluation of how well each outlier detection method performs in accurately tracking the puck's position.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

For all subjective evaluations (user studies #1-#3 related to the smoothing module, overall pipeline, and competitor analysis), we use the Mean Opinion Score (MOS), based on Absolute Category Rating (ACR) for a scale of 1–5, to assess human subjective perception, reflecting how well certain outputs align with viewer experience.

	Model Name	Version	Puck TP	Puck FN
1	Sc-n	Nano	66%	34%
2	Sc-s	Small	75%	26%
3	**Sc-m**	Medium	**77%**	**23%**
4	Sc-l	Large	76%	24%
5	Sc-x	XLarge	78%	22%

TABLE 3: Object detection performance for YOLO models.

C. SCENE DETECTION PERFORMANCE

An exhaustive grid search was conducted on the PySceneDetect model parameters in prediction mode. This involved adjusting key parameters like adaptive threshold, minimum scene length, window width range, and minimum content value. A scoring function was applied to assess the accuracy of PySceneDetect's predictions against the ground truth. This exhaustive search for the optimal parameters in prediction mode yielded three sets of parameter combinations. These combinations achieved the best F1 scores across all five videos and 30,000 frames. As illustrated in Figure 8, the minimum F1 score for the best prediction across all videos was 72%, marking a significant improvement compared to the default parameters of PySceneDetect which was around 57% for the same videos. This enhancement is attributed to the adjustment of parameters to suit the visual characteristics of hockey frames, notably their brightness and color, predominantly featuring a white background.

The optimal combination of hyperparameters that were identified, which can be seen as a red sphere in Figure 9, includes:

luma_only = false;
adaptive_threshold = 1.5;
min_scene_len = 140;
min_content_val = 20;
window_width = [15, 20, 25].

D. OBJECT DETECTION PERFORMANCE

Table 3 delineates the object detection performance of different YOLO model configurations. It highlights that the Sc-m model is the most effective in our pipeline, with a true positive rate of 77% for puck detection with an F1 score of 69%. This high performance is indicated by its ability to accurately identify 77 out of 100 pucks, with only 23 instances where it failed to detect the puck (false negatives), suggesting a favorable balance in its detection capabilities. The confusion matrix in Figure 10 visually represents these detection results for the puck class using the fine-tuned YOLOv8 medium version named `Sc-m` model as in Table 3, illustrating the model's accuracy and error distribution.

The Mean Average Precision (MAP) at the Intersection over Union (IoU) threshold of 0.5 for pucks and all classes is recorded at 0.576. This value indicates a moderate level of accuracy across the model's predictions. In its operation, the model exhibits a trade-off between precision and recall. Initially, the precision is relatively high, indicating strong specificity in early stages. However, as the model strives to capture more true positives, there is a noticeable decline in

(a) Scene 1 (b) Scene 2 (c) Scene 3 (d) Scene 4

FIGURE 7: The first frames from each scene of a 90-second ice hockey video, highlighting scene changes.

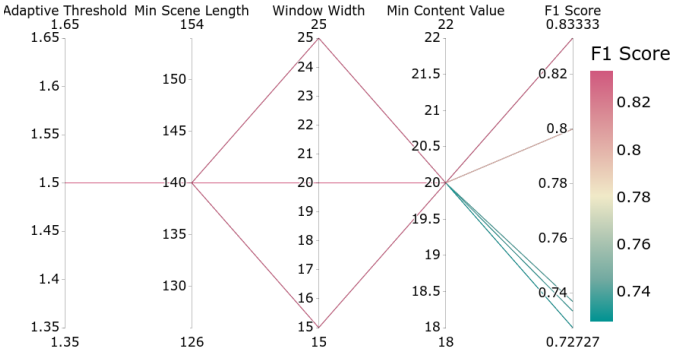


FIGURE 8: Scene detection: discovering optimal hyperparameter combinations.

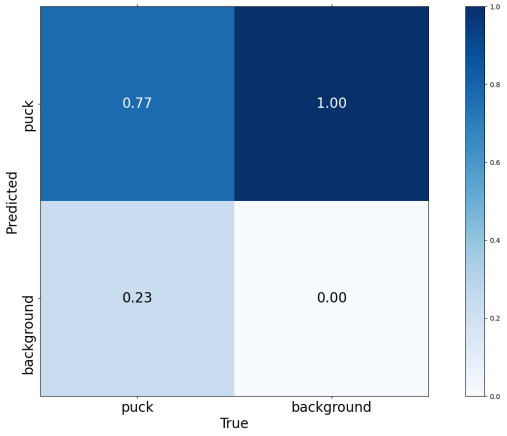


FIGURE 10: Confusion matrix for model Sc-m.

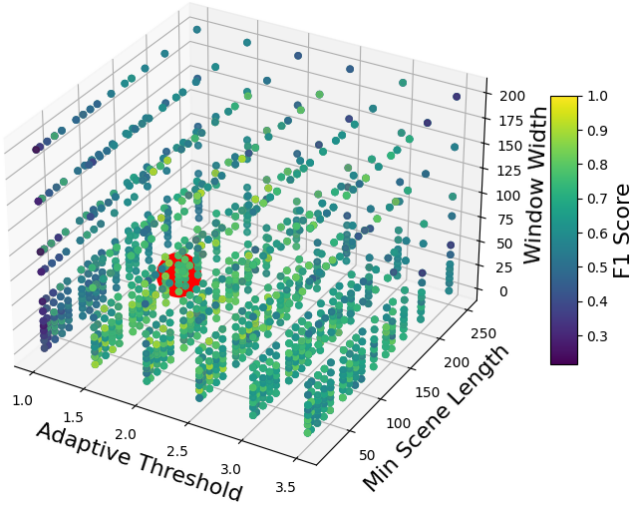


FIGURE 9: 3D visualization of parameter space (red circle = best F1 values).

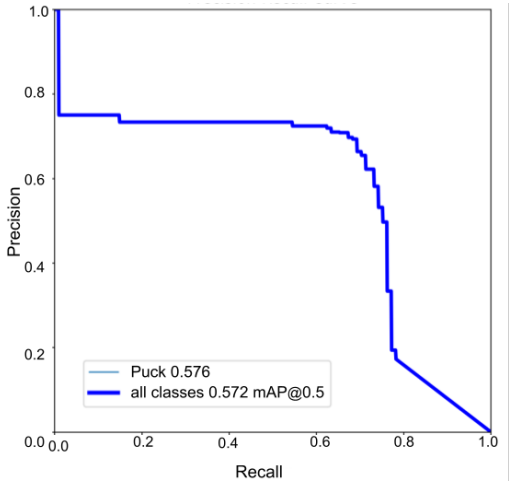


FIGURE 11: Precision-Recal curve for model Sc-m.

precision. This reflects the inherent challenge in balancing precision with recall in object detection tasks, as further illustrated by the Precision-Recall (PR) curve for the puck class using the Sc-m model in Figure 11.

E. OUTLIER DETECTION PERFORMANCE

We compare the performance of different outlier detection methods using our annotated dataset. The MAE is employed to calculate the average absolute difference between the puck positions that remain after applying Z-score, modified Z-score, and IQR outlier detection methods, and the ground truth positions. Table 4 showcases the superior performance of the IQR outlier detection method compared to the other two. Additionally, Figure 12 illustrates that the puck detections remaining after applying the IQR method align more closely with the accurate pucks.

F. SMOOTHING PERFORMANCE

Figure 13 presents a subjective evaluation of video transition smoothness using five hockey videos. Focusing on cases

	Method	MAE
1	Zscore	125.230
2	Modified Z-Score	125.131
3	**IQR**	**95.796**

TABLE 4: Outlier detection performance for different methods.

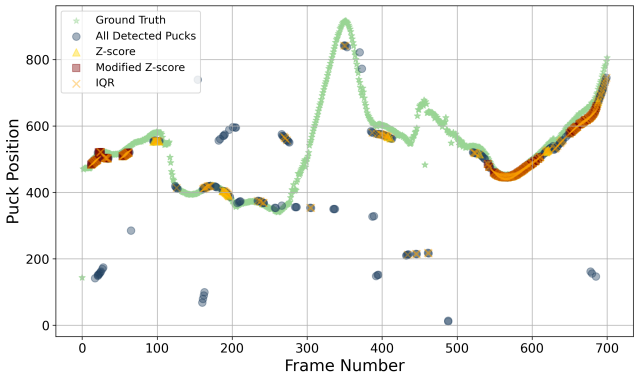


FIGURE 12: Comparative analysis of outlier detection techniques.

where the puck moves across the field, we assessed the smoothness of transitions by modulating the effect with different alternatives: Alternatives 1 and 2 use smooth functions with alpha values of 0.2 and 0.8, while Alternative 3 serves as a baseline without a smooth function. User ratings (Poor=1 to Excellent=5) offer insights into preferred smoothness levels for video editing.

The results from our user study, which involved feedback from 11 participants, demonstrate clear user preferences in video transition smoothness. **Smoothing**, Alpha=0.8, consistently received favorable ratings, indicating a preference for stronger smoothing effects. In contrast, Smoothing, Alpha=0.2 showed moderate acceptance, with a mix of "Good" and "Fair" ratings. Notably, Without Smoothing, was least preferred, often receiving "Bad" and "Poor" ratings. These findings highlight the importance of smooth transitions in video editing and underscore the need for effective smoothing functions to enhance user experience in video editing applications.

V. OVERALL EVALUATION

In this section, we provide an overview of our evaluation process. We analyze the effectiveness, efficiency, and overall impact of the system under review, setting the stage for more detailed discussions in subsequent sections.

A. SYSTEM PERFORMANCE

In video processing pipelines, the quality of the final output is paramount, necessitating high-resolution, clear videos. Simultaneously, it is essential to balance other non-functional requirements like minimal resource consumption and reduced publication latency, ensuring overall efficiency and effectiveness.

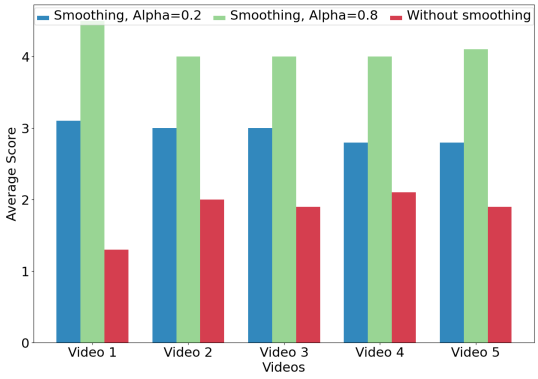


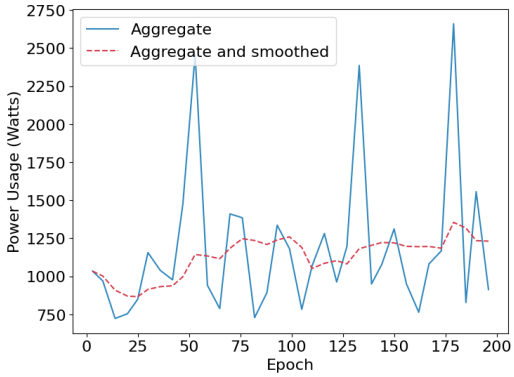
FIGURE 13: [User study #1] Subjective quality ratings for different smoothing alternatives.

1) Model Training

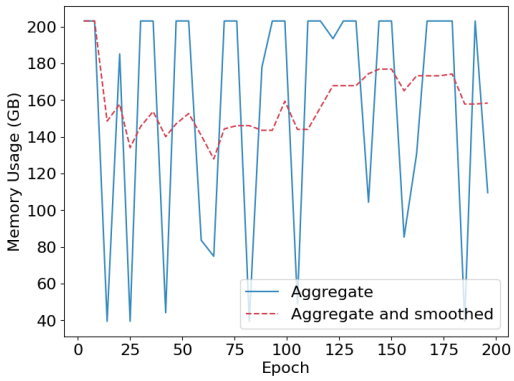
The computational framework for the training of the YOLOv8 object detection model was built on a high-performance computing cluster equipped with eight NVIDIA Tesla V100-SXM3-32GB GPUs. These GPUs, featuring 5,120 CUDA cores and 640 Tensor cores each, are optimized for deep learning tasks. Operating at a base clock of 1,245 MHz and a memory clock of 877 MHz, the GPUs are interconnected via NVLink for efficient multi-GPU scaling. This robust setup was utilized to train the YOLOv8 model, specifically the medium-size variant. The training was carried out with a batch size tailored to the 32GB memory of the GPUs, balancing the need to maximize computational resource usage against the risk of overloading. In the following, we present some results and insights.

Execution time: The training process was completed in a total of 36 minutes and 30 seconds, utilizing the advanced computational resources of the system.

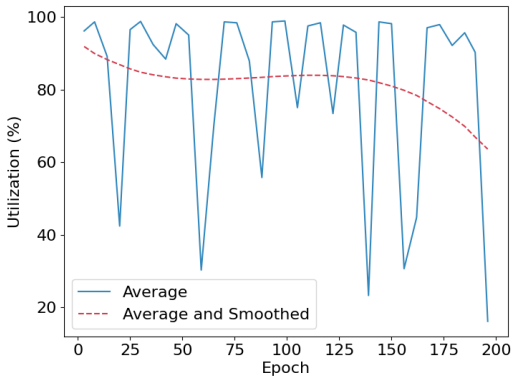
GPU Power Usage: Our analysis of GPU power consumption during model training, as shown in Figure 14a, indicates that the GPUs are operating near their maximum capabilities. With power consumption frequently approaching 250W, or nearly 83% of the V100's specified Thermal Design Power (TDP) of 300W, it is evident that our training process demands intensive computational resources. This high level of GPU utilization, often equated with peak performance in machine learning tasks, is crucial for achieving high throughput and rapid processing speeds. It demonstrates our system's ability to maintain sustained high-performance states, which is pivotal for reducing epoch durations in deep neural network training. The correlation between the substantial power draw and the enhanced training efficacy of our computational framework is unmistakable. Additionally, the consistent operation near the TDP limit indicates robust thermal management within the GPUs, ensuring they avoid throttling and maintain continuous, uninterrupted performance during the training phase. This is in line with our objective of maximizing hard-



(a) GPU power usage



(b) GPU memory usage



(c) GPU utilization

FIGURE 14: Resource consumption for model training (8 GPUs).

ware utility for optimal training efficiency.

GPU memory usage: The memory usage pattern (Figure 14b) indicates that the YOLOv8 model makes extensive use of the available 32GB of HBM2 memory on the Tesla V100s. The variability in memory utilization reflects the data loading and unloading phases inherent to the training epochs.

GPU utilization: Utilization trends (Figure 14c) underscore the episodic nature of GPU demand throughout the training process, with full utilization corresponding to the forward and backward propagation steps, while the troughs may correspond to the evaluation or data preparation stages that are less compute intensive.

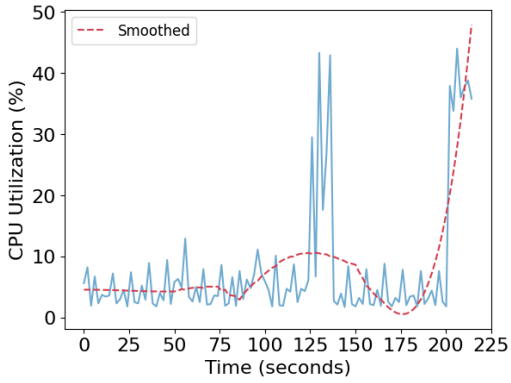
General insights and takeaways: The analysis of resource consumption metrics across the Tesla V100 GPUs reveals a synchronized pattern of utilization, indicative of a balanced computational load across the cluster. This synchronization ensures that no single GPU becomes a bottleneck, contributing to the overall efficiency of the model training process. However, periodic declines in GPU utilization were observed, suggesting the presence of non-computational bottlenecks. These are likely associated with data transfer or preprocessing stages. Refinements in the data pipeline, such as enhanced data transfer protocols or more efficient preprocessing algorithms, could lead to more uniform GPU utilization. Such improvements have the potential to reduce epoch times and improve the throughput of the training pipeline, thereby streamlining the overall workflow.

2) End-to-End Pipeline Execution

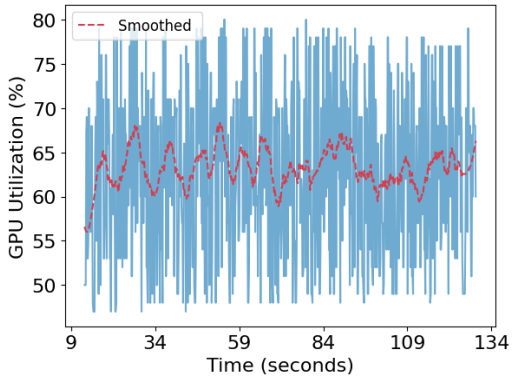
The local deployment was conducted on a system equipped with an NVIDIA Tesla T4 GPU, Driver version 545.23.08, CUDA version 12.3, and a substantial memory size of 16,106 MB. The testing environment included an Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz and an architecture supporting both 32-bit and 64-bit operations. Video input for the tests comprised a 20-second duration clip, running at 50 frames per second (FPS), and a resolution of 1920×1080. To evaluate the impact of the Skip Frame on pipeline execution time, three distinct active configurations were applied (`None`, 13, and 25), with the detection confidence threshold maintained at 0.2, the target aspect ratio set to 1:1, and the output format configured as `mp4`. The pipeline was executed on the same 30-second video with a frame-skipping strategy of 25 frames, as depicted in Figure 16b in red to highlight the allocation of resources for its execution. Below, we present some results and insights.

Execution time: The execution time for each module in the pipeline was measured under all three configurations, with results depicted in Figure 16a. *Constant Execution Time:* Modules such as scene detection, outlier detection, smoothing, cropping, and post-processing demonstrated consistent execution times across all configurations, indicating minimal impact from configuration variations. *Variable Execution Time:* Notably, the object detection module showed significant variance in execution time between the configurations. This variance is hypothesized to stem from the Skip Frame parameter, which was set to 13 and 25 in the second and third configurations, respectively. Adjustments to Skip Frame potentially allowed the object detection algorithm to process fewer frames in these configurations, thereby reducing execution time compared to the first configuration with no frame skipping.

CPU utilization: CPU load was monitored at one-second intervals throughout the pipeline's execution. The object detection module significantly leveraged the GPU, whereas the cropping and post-processing modules predominantly utilized the CPU. The CPU utilization plot, shown in Figure 15a,



(a) CPU



(b) GPU (object detection)

FIGURE 15: Resource utilization for the end-to-end pipeline.

reveals a consistent load with occasional spikes. These spikes can be attributed to the intensive computational demands during the cropping and post-processing phases, highlighting dynamic resource allocation based on task-specific requirements.

GPU utilization (object detection module): The object detection module, as the primary GPU-utilizing component of the pipeline, exhibited variable GPU load correlated to each frame processed. This fluctuation reflects the computational demands imposed by real-time object detection tasks, varying with the density and complexity of objects in each frame. The GPU load plot, as seen in Figure 15b, demonstrates these dynamic shifts in resource allocation. Notably, the scene detection module, while utilizing GPU resources, registered minimal GPU load, often close to 0%. This suggests that the computational requirements for scene detection in our setup are minimal or indicate efficient GPU usage for this specific task. It should be noted that Figure 15b, illustrating the GPU load for object detection, is a subset of Figure 15a. This subset specifically commences from the point in time where GPU load is observed for the object detection module.

Impact of video length on system performance: The introduction of a 30-second video into our test environment provided valuable insights into the scalability of our pipeline. The extended duration, up from the initial 20 seconds, allowed for a more robust assessment of performance stability over time. Initial findings indicate a linear increase in processing time proportional to the video length. This is consistent with expectations, as longer videos naturally demand more computational resources for analysis. However, the efficiency of processing and memory utilization remained consistent, suggesting effective resource management by the system (see Figure 16b).

Comparative analysis of CPU and GPU performance: A key aspect of our study was comparing the performance of the system when running solely on the CPU versus the GPU-enhanced configuration. As anticipated, the GPU-accelerated environment demonstrated markedly superior performance, particularly in terms of reduced execution times. This is attributed to the parallel processing capabilities of the NVIDIA Tesla T4 GPU, which excels in handling the intensive computational demands of video processing tasks. In contrast, the CPU-only configuration, despite its robust capabilities (Intel(R) Xeon(R) Gold 6130 CPU), exhibited longer processing times, underscoring the GPU's pivotal role in accelerating video analysis tasks (see Figure 16c).

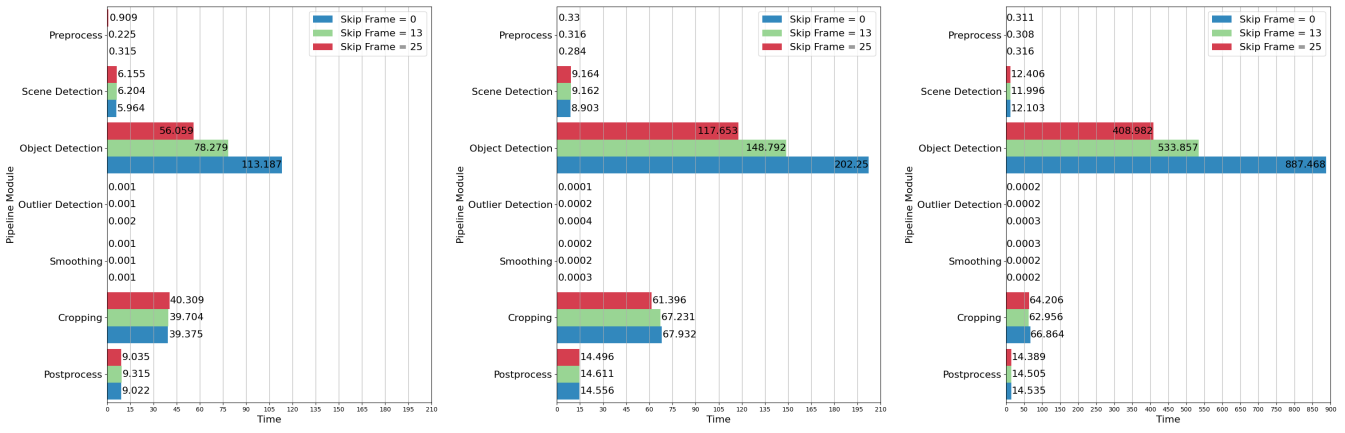
Observations on CPU processing power threshold: The addition of a CPU-only test scenario brought to light the threshold limits of CPU processing in our pipeline. Notably, when dealing with high-definition video inputs and complex detection algorithms, the CPU's processing time increased significantly, more so than with the GPU. This observation is particularly pertinent when considering the handling of high-frequency data, such as in our 50 FPS video. While the CPU managed to maintain functional integrity without system crashes, the efficiency drop was notable. These findings underscore the importance of GPU acceleration in scenarios demanding high-speed data processing and complex computations, such as in advanced video analysis systems. The variable execution times in the object detection module were further accentuated in the CPU-only configuration, especially with the longer 30-second video. This reinforces the hypothesis that the Skip Frame parameter's impact is more pronounced in environments with limited parallel processing capabilities, such as a CPU-only setup. Future investigations may explore the scalability of different CPU architectures and their thresholds to handle video analysis tasks without GPU support.

B. SUBJECTIVE USER EXPERIENCE

We conducted a user study to assess the user QoE of the SmartCrop-H pipeline. More specifically, we investigated six different cropping alternatives (as listed in Table 5, where type 6 corresponds to the SmartCrop-H with full functionality), on four videos (all with an original aspect ratio of 16:9, where two of them were cropped to 9:16, and two were cropped to 1:1), to compare viewer experience.

1) Experimental Setup

We used an online survey to retrieve responses from participants in a remote-study fashion [52], [53]. The survey consists



(a) 20-Second Video on a GPU (b) 30-Second Video on a GPU (c) 30-Second Video on a CPU

FIGURE 16: Runtime per module in local deployment, with and without Skip Frame for 20 and 30 seconds video (note: different x-axis).

	Center	Description	Outlier Detection	Smoothing
1	frame	static no padding	✗	✗
2	frame	static w/black padding to 16:9	✗	✗
3	puck	use last detected puck position	✗	✗
4	puck	w/smoothing	✗	✔
5	puck	w/outlier detection	✔	✗
6	puck	w/outlier detection & smoothing	✔	✔

TABLE 5: Cropping methods used in subjective study #2.

of six pages: (1) Introduction and pre-questionnaire, (2-5) One questionnaire page per case, and (6) Post-questionnaire. In the Introduction, participants were asked to complete the survey on a mobile phone, as this provides a more realistic setting for our evaluation [54], and to view the clips in full-screen mode. Participants first viewed the original video in the 16:9 aspect ratio before evaluating each cropping alternative using a 5-point Absolute Category Rating (ACR) scale, as recommended by ITU-T P.910 [55]. Three questions were posed to assess various aspects: overall QoE (*"How was your overall experience with this video clip?"*), the smoothness of the video (*"How was the smoothness of the window transitions in the video clip?"*), and the video's ability to capture the essence of the original content (*"How well do you think the cropped video captures the essence of the original video?"*).

2) Participant Details

In total, we recruited 26 participants: 5 females, 20 males, no other, and 1 preferring not to disclose gender. The age of the participants ranged from 20 to 60, with a mean of 31.8 and a standard deviation of 8.08. All participants were active on social media. Regarding daily usage, 62% reported spending less than 30 minutes, 19% reported 30 minutes to 1 hour, 12% reported 1-2 hours, and 0.08% reported 2-4 hours. On a scale of 1 to 5, participants indicated that they have experience in video editing with an average score of 2.08.

3) Results

Figure 17 presents the results of our second user study. We have performed ANOVA analysis, where three statistical measures are essential:

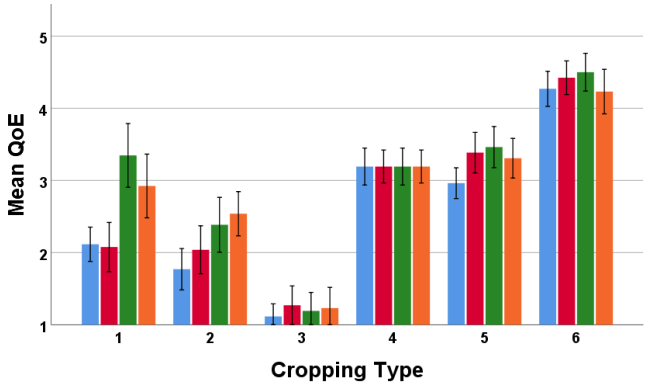
- **F-value**: Indicates the variance ratio between and within groups. A higher F-value suggests significant differences among group means, implying that different cropping styles notably affect participant ratings in our study.
- **p-value**: Measures the probability of observing the results under the null hypothesis. A p-value below 0.05 typically indicates statistical significance. Our p-value less than .001 strongly suggests the observed differences in ratings are not due to chance.
- **Partial** η^2: Represents effect size, showing the proportion of variance explained by a factor. A higher Partial η^2 indicates that the factor (cropping style) substantially impacts the observed outcome (QoE).

These measures collectively inform the significance and magnitude of the effect of cropping styles on the QoE in our study.

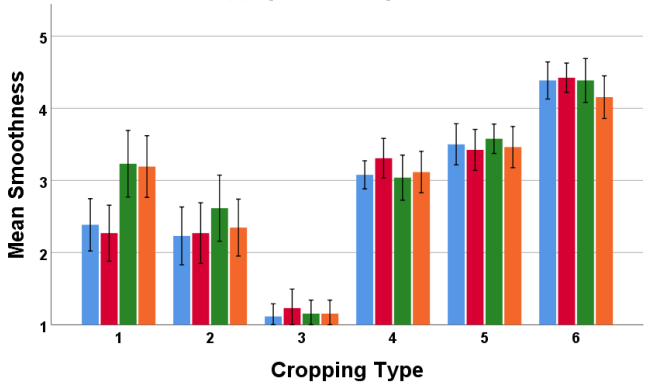
Quality of experience (QoE)

To evaluate the influence of cropping styles on viewer perception of video content, a series of One-Way Repeated Measures ANOVAs [56] were employed. These analyses focused on determining the effect of six different cropping methods (detailed in Table 5) on participant ratings for four different video scenarios.

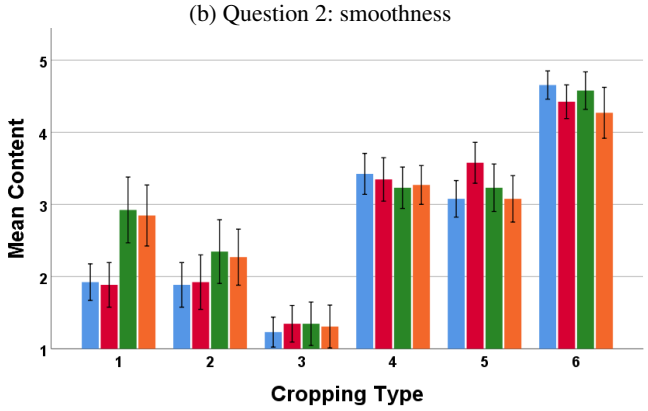
As shown in Figure 17a, for Video 1, the analysis revealed a significant effect of cropping type on participant ratings in the first row in Table 6. Descriptive statistics indicated that cropping type 6 received the highest mean rating, suggesting



(a) Question 1: QoE



(b) Question 2: smoothness



(c) Question 3: content

FIGURE 17: [User study #2] Subjective quality ratings with 95% confidence intervals, for Video 1 (blue) and Video 2 (red) with a target aspect ratio of 9:16, Video 3 (green) and Video 4 (orange) with a target aspect ratio of 1:1.

it was the most favorably perceived cropping style among participants.

In the case of Video 2 in the second row in Table 6, there was a significant effect of cropping type on participant ratings. The highest mean rating was observed for cropping type 6, indicating it as the most preferred cropping style.

For Video 3 in QoE metric in Table 6, the analysis showed a significant effect of cropping type on participant ratings. Cropping type 6 emerged as the most preferred type with the highest mean rating.

Regarding Video 4 in QoE metric in Table 6, a significant effect of cropping type on the ratings was observed. Cropping type 6 was rated the highest among all types, suggesting it is the most favorable cropping style.

Smoothness

A repeated measures ANOVA was performed to investigate the effects of six varied cropping techniques on perceived smoothness across four different scenarios. The results from this analysis revealed a significant influence of the cropping type on the perceived smoothness in every scenario as shown in Figure 17b and in Smoothness metric in Table 6.

For Video 1, the analysis revealed a significant effect of cropping type on video ratings. Cropping type 6 (SmartCrop-H) was rated highest among all types, indicating a strong preference for this style. This significant finding highlights the influence of cropping style on viewer perceptions, with Cropping type 6 distinctly standing out as the most favored by participants.

In the case of Video 2, a One-Way Repeated Measures ANOVA demonstrated a significant effect of cropping type on video content ratings. The results showed that Cropping type 6 (SmartCrop-H) was rated the highest mean rating among all cropping types. This result emphasizes the substantial influence of cropping style on viewer perceptions, highlighting a strong preference for Cropping type 6 in this context.

For Video 3, a significant effect of cropping type on video ratings was identified. Cropping type 6 (SmartCrop-H) emerged as the most favored cropping style, receiving the highest mean rating. This finding underscores the significant influence of cropping type on viewer preferences, with a pronounced preference for cropping type 6 in the assessed dataset.

Regarding Video 4, the analysis indicated a significant effect of cropping type on video content ratings. Among the six cropping types evaluated, Cropping type 6 (SmartCrop-H) received the highest mean rating. This outcome highlights the significant impact of cropping style on viewer perceptions, with cropping type 6 emerging as the most preferred by participants.

Capturing the Essence of the Original Video (Content)

To evaluate the effects of six distinct cropping methods on preserving the essence of the original video content, a one-way repeated measures ANOVA was carried out across four scenarios. The outcomes of this analysis are illustrated in Figure 17c in Content metric in Table 6.

For Video 1, a One-Way Repeated Measures ANOVA revealed a significant effect of cropping type on the ratings of video content. Among the evaluated cropping types, cropping type 6 was rated highest mean. This finding emphasizes the substantial influence of cropping type on viewer preferences, highlighting a distinct preference for cropping type 6.

In the case of Video 2, the analysis showed a significant effect of cropping type on video ratings. Cropping type 6

Metric	Video #	Aspect Ratio	Mean	SD	F-value	p-value	Partial η^2
QoE	1	9:16	4.2692	0.60383	92.899	$< .000$	.957
QoE	2	9:16	4.4231	0.57779	53.945	$< .000$	.928
QoE	3	1:1	4.5000	0.64807	46.014	$< .000$	.916
QoE	4	1:1	4.2308	0.76460	29.516	$< .000$	.875
Smoothness	1	9:16	4.3846	0.63730	112.051	$< .000$	.964
Smoothness	2	9:16	4.4231	0.50383	72.028	$< .000$	.945
Smoothness	3	1:1	4.3846	0.75243	100.258	$< .000$	.960
Smoothness	4	1:1	4.1538	0.73170	47.933	$< .000$	.919
Content	1	9:16	4.6538	0.48516	96.253	$< .000$	.958
Content	2	9:16	4.6538	0.5779	91.917	$< .000$	.786
Content	3	1:1	4.5769	0.64331	41.102	$< .000$	.907
Content	4	1:1	4.5769	0.87442	30.677	$< .000$	.880

TABLE 6: [User study #2] Summary of the ANOVA analysis for the QoE, smoothness, and content metrics for SmartCrop-H across four videos.

was rated the highest among all types, indicating a strong preference for this particular cropping style with a mean rating. This result underscores the considerable impact of cropping style on viewer perceptions, with a clear inclination toward cropping type 6.

For Video 3, a One-Way Repeated Measures ANOVA indicated a significant effect of cropping type on video ratings. Cropping type 6 emerged as the most preferred type with the highest mean and standard deviation rating, reflecting a strong preference for this cropping style among the participants.

Regarding Video 4, the analysis demonstrated a significant effect of cropping type on the video content ratings. Cropping type 6 was rated the highest mean, suggesting it as the most favored cropping style. This result highlights the significant impact of cropping style on viewer preferences, with a pronounced inclination towards cropping type 6 in the evaluated dataset.

4) Post-questionnaire

In the follow-up questionnaire, participants evaluated the significance of four key elements in video viewing on a 1 to 5 scale, with 1 denoting "Not at all important" and 5 indicating "Extremely important." The evaluated elements included *smooth window movement*, *visibility of relevant players*, *visibility of the puck*, and *overall video quality*. The collective feedback from all 26 participants can be summarized as follows.

- **Smooth window movement:** A mean score of 3.42 was observed, with the most frequently chosen ratings being 3 (*Important*).
- **Always seeing relevant players:** This factor achieved a mean rating of 3.35, with the most common score being 3 (*Important*).
- **Always seeing the puck:** This element scored an average of 3.28, with the predominant rating being 3 (*Important*).

	Description	Automatic
1	Static cropping (frame-centered)	✔
2	SmartCrop-H	✔
3	Final Cut Pro Auto Reframe	✔
4	Adobe Premiere Pro Auto Reframe	✔
5	Manual cropping with Adobe Premiere Pro	✖

TABLE 7: Cropping methods used in subjective study #3.

- **Video quality:** Garnered an average rating of 3.54, with the most frequently selected score being 3 (*Important*).

The survey results suggest that participants regard all four aspects - smooth window movement, visibility of relevant players and the puck, and video quality - as important, albeit with a slight inclination towards the visibility elements over technical quality, indicating a preference for content over video technicalities.

C. COMPETITOR ANALYSIS

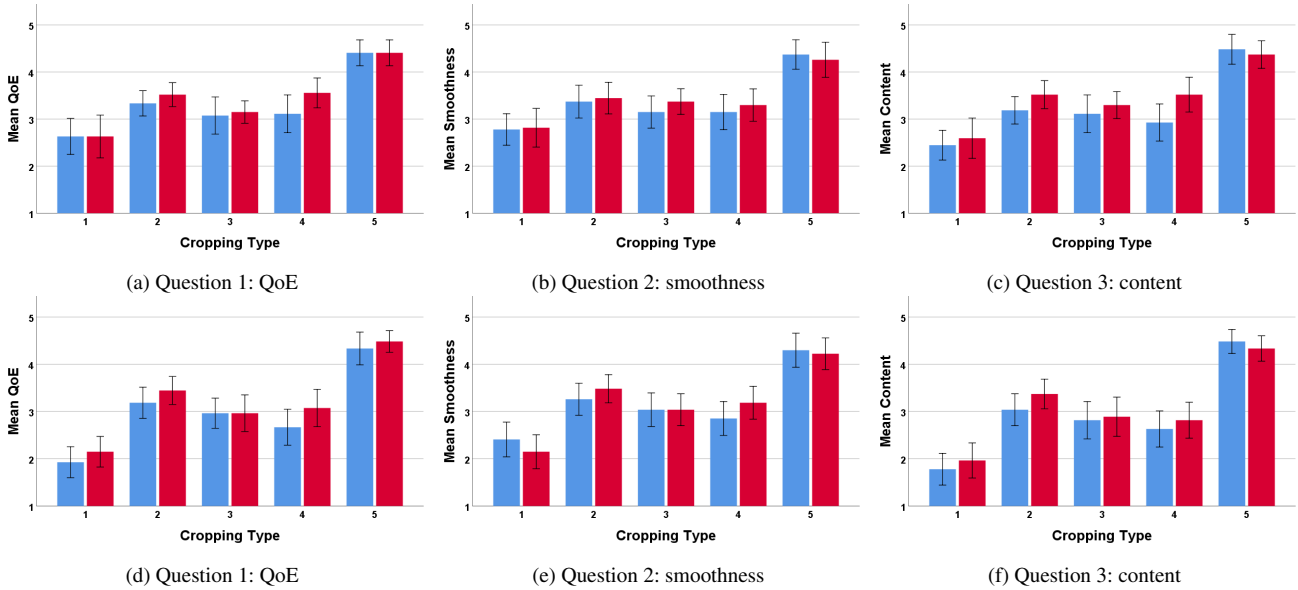
In order to undertake a competitor analysis for the cropping operation, we conducted a user study comparing the cropping methods presented in Table 7. This study involved examining two distinct videos, each cropped using the five different cropping methods. The first alternative involves static frame-centered cropping. The second represents the output from our SmartCrop-H pipeline. The third alternative is the *Final Cut Pro Auto Reframe*, an automated cropping feature within Final Cut Pro. The fourth, *Adobe Premiere Pro Auto Reframe*, operates similarly but within the Adobe Premiere Pro environment. Finally, the fifth alternative is manual cropping, wherein a human video producer manually crops the video using the Adobe Premiere Pro software. The videos were evaluated across two aspect ratios, namely 1:1 and 9:16.

1) Experimental Setup

We conducted an online survey, gathering responses from participants in a collaborative manner. The survey structure mirrored that of section V-B1, where participants were tasked with completing it using an iPhone. The choice of using iPhones was due to a compatibility issue: videos could not be displayed in their full size when streamed from YouTube on Google Forms on Android devices. Ensuring that participants could view the video in full-screen was only feasible on iPhones, providing a consistent and optimal viewing experience for all respondents.

2) Participant Details

In total, we recruited 27 participants: 7 females, 20 males. The ages of the participants ranged from 17 to 40, with a mean of 31.7 and a standard deviation of 7.74. All participants were active on social media. Regarding daily usage, 29.6% reported spending less than 30 minutes, 18.5% reported 30 minutes to 1 hour, 29.6% reported 1-2 hours, 14.8% reported 2-4 hours, and 7.4% reported more than 4 hours. On a scale of 1 to 5, participants indicated that they have experience in video editing with an average score of 1.96.



(a) Question 1: QoE (b) Question 2: smoothness (c) Question 3: content

(d) Question 1: QoE (e) Question 2: smoothness (f) Question 3: content

FIGURE 18: [User study #3] Subjective quality ratings with 95% confidence intervals, for Video 1 (blue) and Video 2 (red) with a target aspect ratio of 1:1 (a-c) and 9:16 (d-f).

3) Results

To analyze the results, we used the repeated measures ANOVA which compares means across one or more variables that are based on repeated observations. The one-way repeated measures ANOVA results are summarized in Table 8.

Metric	*Video #*	*Aspect Ratio*	*Mean*	*SD*	*F-value*	*p-value*	*Partial η^2*
QoE	1	1:1	3.4444	0.75107	29.932	$< .001$	0.535
QoE	2	1:1	3.5185	0.64273	25.344	$< .001$	0.494
QoE	1	9:16	3.1852	0.83376	37.951	$< .001$	0.593
QoE	2	9:16	3.4444	0.75107	29.932	$< .001$	0.535
Smoothness	1	1:1	3.3704	0.88353	14.779	$< .001$	0.362
Smoothness	2	1:1	3.4444	0.84732	11.596	$< .001$	0.308
Smoothness	1	9:16	3.2593	0.85901	19.110	$< .001$	0.424
Smoothness	2	9:16	3.4815	0.75296	23.061	$< .001$	0.470
Content	1	1:1	3.1852	0.73574	25.250	$< .001$	0.493
Content	2	1:1	3.5185	0.75296	19.545	$< .001$	0.429
Content	1	9:16	3.0370	0.85402	48.494	$< .001$	0.651
Content	2	9:16	3.3704	0.79169	24.245	$< .001$	0.483

TABLE 8: [User study #3] Summary of the ANOVA analysis for the QoE, smoothness, and content metrics for SmartCrop-H across two videos and two target aspect ratios.

Quality of experience (QoE)

The repeated measures ANOVA assessed the impact of five cropping methods on QoE across four scenarios (Table 7). The MOS values for a 1:1 aspect ratio are shown in Figure 18a and for a 9:16 aspect ratio in Figure 18d.

For Video 1 with a 1:1 aspect ratio, SmartCrop-H showed a notable mean performance score of 3.44. This result was significantly better than the Static Crop method and slightly superior to the automated reframing techniques of Final Cut Pro and Adobe Premiere Pro. However, it did not reach the excellence of a manually edited crop using Adobe Premiere Pro, which had the highest mean score. This indicates SmartCrop-H's effectiveness over traditional and less sophisticated methods, while also suggesting room for improvement to match the precision of manual cropping.

In the 9:16 aspect ratio for Video 1, SmartCrop-H demonstrated robustness with a mean score of 3.1852. Again, it significantly outperformed the Static Crop and exhibited better results compared to the automated methods, but it did not surpass the Manual Crop. This trend was also observed in Video 2 across both aspect ratios, where SmartCrop-H consistently showed improvement over basic and automated cropping methods but fell short of the high standards set by manual cropping.

Smoothness

The analysis of video smoothness indicated a significant impact of cropping type on both videos in each aspect ratio, as shown in (Figure 18b and Figure 18e). SmartCrop-H demonstrated effective performance, especially in the 1:1 aspect ratio for Video 1, with a mean score of 3.3704. It notably outperformed the Static Crop and showed better results than the automated methods. However, as expected, it was not on par with the Manual Crop's performance. This pattern was consistent across all scenarios, underscoring SmartCrop-H's efficiency in automated editing yet highlighting the superiority of manual techniques.

Content

In capturing the original content, SmartCrop-H's performance in the 9:16 aspect ratio was particularly noteworthy (Figure 18f). For Video 1, it had a mean performance of 3.2593, significantly better than the Static Crop. Despite sur-

passing other automated methods, it did not reach the level of the Manual Crop. This trend was similar for Video 2, where SmartCrop-H effectively outperformed basic and some automated methods but did not match the quality of manual cropping.

Across all scenarios, SmartCrop-H showed significant superiority over Static Crop and slight improvements over Final Cut Pro Auto Reframe (Type 3) and Adobe Premiere Pro Auto Reframe (Type 4). However, it was consistently outperformed by Manual Crop (Type 5), which had the highest mean score (4.4815). These results suggest SmartCrop-H's advancement over basic and less sophisticated methods, its marginal superiority over other automated techniques, and the potential for further development to reach the standards of manual cropping, as exemplified by the Manual Crop results.

4) Post-questionnaire

In the subsequent survey, participants rated the importance of four critical aspects of video watching on a scale from 1 to 5, where 1 represented "Not at all important" and 5 signified "Extremely important." These aspects were *Smoothness of window movement*, *Always seeing relevant players*, *Always seeing the puck*, and *Video quality*. Summarized responses from the 27 participants were as follows:

- **Smooth window movement:** With an average score of 2.7 and a mode of 3 ("Important"), it suggests this feature is considered moderately important, but not a top priority.
- **Always seeing relevant players:** This aspect's high average score of 4 and a mode of 5 ("Extremely important") clearly indicates it is a crucial element for viewers.
- **Always seeing the puck:** An average score of 3.37 and a mode of 3 ("Important") reflect a general consensus on its importance, but not as vital as seeing relevant players.
- **Video quality:** The average score of 3.19 and a mode of 3 ("Important") indicate that while video quality is important, it is less critical compared to the visibility of players and the puck.

These results indicate that participants place considerable importance on all four aspects, with smoothness and video quality shown to be slightly less critical than the visibility of players and the puck (indicating prioritization of content over video quality).

5) Saving Human Resources - Time Spent for Cropping

One of the primary objectives of our project is to streamline the video editing process, with a specific focus on reducing the time and human resources required for publishing content across various platforms. Our experiment also involved a comparative analysis of manual and AI-assisted video cropping techniques regarding their execution time. Initially, we manually cropped an 18-second video highlight using Adobe Premier Pro. This process, including the setting of specific cropping points for each frame, took us approximately 240 seconds. Notably, this time frame did not include the additional time required for the final video export, which encompasses transcoding and file writing.

In contrast, when we applied our SmartCrop-H to the same 18-second video clip, the results were significantly more efficient. Our pipeline utilizes AI to analyze scenes, detect key elements in the footage (such as players and the puck in our case), track objects in each frame, interpolate, smooth the transitions between frames, and then calculate the optimal cropping points. The entire process, from detecting these elements to producing a cropped video ready for publishing, took merely 62 seconds.

Similarly, we observed that the Auto Reframe feature in Adobe Premiere Pro, and a comparable feature in Final Cut, both of which employ AI, significantly expedite the reframing process. These tools analyze videos for color and saliency to determine how best to reframe them. While it is difficult to precisely measure the time taken by these black-box systems, the process is notably quicker than manual methods. The user simply drags the auto-raframe effect onto the video, after which the system analyzes color and saliency. The reframed views become available shortly afterward. If desired, this can then be exported, initiating a crop, transcoding, and writing to file. To utilize these features effectively, users require a robust platform with adequate hardware. For Adobe Premiere Pro and Final Cut, they recommend a system with at least 4GB of VRAM, 16GB of RAM, and a multi-core processor (preferably 6-core or higher). These specifications ensure smooth operation and quick processing times, enabling these AI-driven tools to function at their full potential.

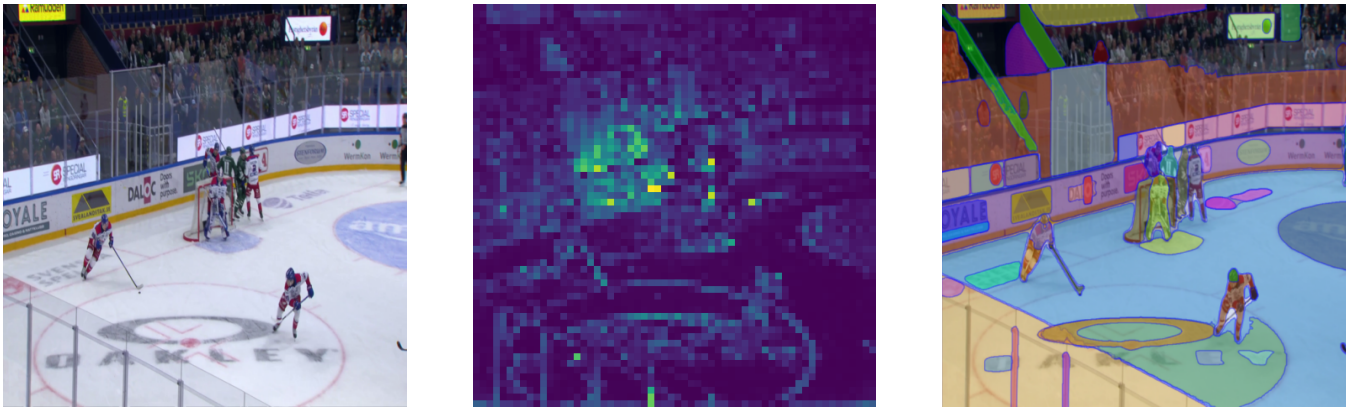
Our findings suggest that AI solutions offer a significant time-saving advantage in adjusting video aspect ratios for various publishing platforms.

VI. DISCUSSION AND FUTURE WORK

This study introduces a novel framework designed specifically for adapting hockey videos to various aspect ratios, primarily for social media distribution. It effectively integrates advanced techniques such as scene analysis, object and outlier detection, and smoothing algorithms to optimize the QoE for viewers across different aspect ratios. This framework's efficacy has been rigorously validated through objective and subjective means. Significantly, as highlighted in Section II, our research emphasizes the importance of domain-specific insights in designing automated production pipelines, particularly in sports where unique optimization strategies are vital.

We examined the efficiency and performance of SmartCrop-H in comparison to manual and automated video editing methods. Our results show that SmartCrop-H outperforms basic methods like Static Crop and exhibits marginal improvements over other automated tools, such as Final Cut Pro's and Adobe Premiere Pro's Auto Reframe methods. However, it is slightly less effective than the high-quality results achieved through manual editing with Manual Crop.

We also examined the time efficiency of SmartCrop-H in video editing, particularly for sports content. While manually editing an 18-second ice hockey video in Adobe Premiere Pro



(a) Ice hockey frame (b) DINO attention visualization (c) Segment Anything visualization

FIGURE 19: Set of images displaying various analysis techniques on a sample ice hockey frame.

(a) Game frame 1 (b) Game frame 2 (c) Game frame 3

(d) Normalized view 1 (e) Normalized view 2 (f) Normalized view 3

FIGURE 20: Comparison of original game frames and their 2D normalized views post-homography.

took about 4 minutes, this timeframe may seem manageable for a single clip but becomes significantly more challenging when dealing with multiple, much longer videos. The scalability issue is crucial, especially for media teams in clubs or leagues that regularly process a large volume of game footage. The manual approach, while initially seeming quick, becomes a daunting, time-consuming task when multiplied across numerous events.

In stark contrast, SmartCrop-H completed the same editing task in just 62 seconds. This substantial reduction in editing time is not just a matter of convenience but a critical efficiency improvement for scenarios involving high content volumes. Furthermore, SmartCrop-H offers an integrated system that accelerates various editing processes and enables the fully automated publishing of videos immediately after a game, which drastically cuts down on manual work. Additionally, SmartCrop-H's features, such as the ability to track multiple objects in a frame, cater to a broader range of viewer interests. This is a significant advantage over Adobe's Auto Reframe feature, which mainly focuses on actions within a frame. This approach by SmartCrop-H enhances the overall viewing experience for different types of clips, showcasing its adaptability and potential in providing tailored video editing solutions. This time efficiency of SmartCrop-H is crucial in fast-paced environments like news and sports media, where quick turnaround is essential. Although it does not match the precision of manual editing, SmartCrop-H provides a valuable balance between speed and quality, making it a useful tool for various applications. These findings suggest the potential of SmartCrop-H to streamline video editing processes, although there is still room for improvement to match the quality of manual editing fully.

The next phase of our research envisages the incorporation of **emerging technologies**. Despite the promise these technologies hold, several technical and logistical challenges have delayed their implementation. The integration of the Segment

Anything Model (SAM) [57], an advanced tool for object segmentation within a frame, is a case in point. Adapting SAM to our specific requirements is a complex endeavor, and ensuring its effectiveness and accuracy in the dynamic environment of a hockey game demands comprehensive testing and calibration (Figure 19c).

Similarly, the inclusion of DINO [58] (self-distillation with no labels) for visual attention modeling introduces its unique set of challenges. Tailoring DINO's AI-driven attention modeling to the specifics of our framework requires a nuanced understanding of its underlying algorithms and the development of a custom interface (Figure 19b). The integration of these technologies, while challenging, is seen as a crucial step in advancing our framework's capabilities.

Furthermore, the expansion of the ROI paradigm to encompass player positions and key game moments signifies a critical shift from traditional puck-centric analyses in ice hockey footage. This shift is essential for optimizing the viewer experience on devices with varying aspect ratios, necessitating precise cropping of each frame to maintain the saliency of the video content. Special attention to game events such as face-offs, power plays, and breakaways is imperative, as these moments where player positions and actions become focal, must be clearly visible, especially on smaller screens. The challenge involves dynamically adjusting the cropping to ensure these pivotal moments are captured effectively.

Our initial attempts at automating the extraction of ROI from ice hockey game footage for social media were based on a multi-faceted approach that included homography, object detection, and the application of ice hockey game rules. Despite the theoretical promise of this approach, practical challenges emerged, particularly due to the fast-paced nature of ice hockey. Homography, aimed at transforming video frames into a normalized 2D view of the rink, was combined with object detection techniques to identify game elements and the application of specific ice hockey rules (Figure 20). However, the rapid pace, unpredictable movements, and the small, fast-moving puck presented significant difficulties in maintaining accuracy and effectiveness in ROI identification.

Despite these challenges, the potential benefits of homography in understanding active areas within a frame are evident. This insight is crucial for developing a more sophisticated approach to video cropping focused on active game areas. Consequently, our future efforts will be directed toward refining the use of homography in combination with enhanced object detection algorithms and a deeper integration of ice hockey rules. This multifaceted approach aims to develop a robust system capable of effectively handling the fast-paced and dynamic nature of ice hockey, thereby improving the automated production and adaptation of sports footage for diverse media platforms.

VII. ETHICAL CONSIDERATIONS AND ENVIRONMENTAL IMPACT

The development and deployment of AI-driven video editing tools such as SmartCrop-H raise several ethical considerations that need discussion. Automation of video editing tasks has the potential to impact **employment** in the media industry, potentially leading to job displacement for video editors specializing in content retargeting for social media. However, these tools may also create new opportunities for content creators to focus on higher-level creative tasks.

There are also concerns about **content integrity**, as automated cropping decisions can inadvertently alter the intended message or emotional impact of the original footage. As with any AI system, there is a potential for bias in the decision-making process, which could manifest itself as favoring certain types of play or players, potentially skewing the representation of the game. The examination of this potential bias can go hand-in-hand with the technical investigation of alternative cropping windows that center around a different POI than the hockey puck (the expansion of the ROI paradigm as discussed earlier).

The **environmental impact** of AI systems is another important consideration that needs to be taken into account. Although we do not provide specific energy consumption information, the use of GPU acceleration in our pipeline suggests potential for improved efficiency compared to CPU-only processes. Our analysis shows that SmartCrop-H, running on a system with an NVIDIA Tesla T4 GPU, completes video processing tasks more quickly than CPU-only configurations, which could translate to energy savings in high-volume scenarios which are typical in sports broadcasting. Future work can undertake more detailed comparisons of energy consumption between SmartCrop-H, manual editing processes, and other automated systems, as well as explore the use of more energy-efficient algorithms and hardware, and the potential for edge computing to further reduce the environmental footprint of AI-driven video processing.

VIII. CONCLUSION

This study introduces the SmartCrop-H pipeline, designed to enhance the viewing of ice hockey videos in various aspect ratios. This system incorporates object detection, scene detection, and outlier detection to identify POIs effectively. A smoothing algorithm further refines the video cropping process. Our object detection model targeting the ice hockey puck showed superior performance compared to other models, and our smoothing module was shown to significantly improve cropping window movement and transition smoothness. Subjective studies have demonstrated the effectiveness of SmartCrop-H in improving viewer QoE, especially for the challenging 9:16 aspect ratio. Overall, SmartCrop-H marks a significant advancement in sports video processing, offering efficient adaptation of video content for varied aspect ratios, enhancing the viewing experience of digital sports media. Future work includes refining the pipeline, integrating emerging technologies, and a broader subjective evaluation to understand viewer preferences across different sports.

CONFLICT OF INTEREST

Authors Dag Johansen and Pål Halvorsen own shares in the company Forzasys AS, which develops AI solutions for sports. Data from Forzasys systems were used for the experiments in this work. Results from the research may eventually be used by Forzasys for content re-targeting. Otherwise, the authors declare no competing interests.

ACKNOWLEDGMENT

This research was funded by the Research Council of Norway, project number 346671 (AI-Storyteller).

REFERENCES

[1] M. H. Sarkhoosh, S. M. M. Dorcheh, S. Gautam, C. Midoglu, S. S. Sabet, and P. Halvorsen, "Soccer on social media," *arXiv preprint arXiv:2310.12328*, 2023.
[2] C. Midoglu, S. S. Sabet, M. H. Sarkhoosh, M. Majidi, S. Gautam, H. M. Solberg, T. Kupka, and P. Halvorsen, "AI-based sports highlight generation for social media," in *Proc. 3rd Mile-High Video Conf. (MHV)*, 2024.
[3] Adobe, "Professional video editing software | Adobe Premiere Pro." https://www.adobe.com/products/premiere.html, 2024.
[4] Apple, "Final Cut Pro for Mac - Apple." https://www.apple.com/final-cut-pro/, 2024.
[5] M. H. Sarkhoosh, D. Sayed Mohammad Majidi, C. Midoglu, S. S. Sabet, T. Kupka, M. A. Riegler, D. Johansen, and P. Halvorsen, "AI-Based cropping of soccer videos for different social media representations," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, 2024.
[6] S. M. M. Dorcheh, M. H. Sarkhoosh, C. Midoglu, S. S. Sabet, T. Kupka, M. A. Riegler, D. Johansen, and P. Halvorsen, "SmartCrop: AI-Based cropping of soccer videos," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, 2023.
[7] H. Nam, D. Park, and K. Jeon, "Jitter-robust video re-targeting with kalman filter and attention saliency fusion network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 858–862, 2020.
[8] H.-S. Lee, G. Bae, S.-I. Cho, Y.-H. Kim, and S. Kang, "Smartgrid: video retargeting with spatiotemporal grid optimization," *IEEE Access*, vol. 7, pp. 127564–127579, 2019.
[9] K.-K. Rachavarapu, M. Kumar, V. Gandhi, and R. Subramanian, "Watch to edit: video retargeting using gaze," *Comput. Graph. Forum*, vol. 37, pp. 205–215, 2018.
[10] E. Jain, Y. Sheikh, A. Shamir, and J. Hodgins, "Gaze-driven video re-editing," *ACM Trans. Graph. (TOG)*, vol. 34, no. 2, pp. 1–12, 2015.
[11] T. Deselaers, P. Dreuw, and H. Ney, "Pan zoom scan – time-coherent trained automatic video cropping," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1–8, 2008.
[12] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," in *Proc. ACM Int. Conf. Multimedia (MM)*, pp. 241–250, 2006.
[13] H. Kaur, S. Kour, and D. Sen, "Video retargeting through spatio-temporal seam carving using kalman filter," *IET Image Process.*, vol. 13, no. 11, pp. 1862–1871, 2019.
[14] S. Wang, Z. Tang, W. Dong, and J. Yao, "Multi-operator video retargeting method based on improved seam carving," in *Proc. IEEE Inf. Technol. Mechatro. Eng. Conf. (ITOEC)*, pp. 1609–1614, 2020.
[15] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," in *Proc. ACM SIGGRAPH*, pp. 1–9, 2010.
[16] S. Kopf, T. Haenselmann, J. Kiess, B. Guthier, and W. Effelsberg, "Algorithms for video retargeting," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 819–861, 2011.
[17] K. Apostolidis and V. Mezaris, "A fast smart-cropping method and dataset for video retargeting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 2618–2622, 2021.
[18] X. Yang, "Where is the puck? tiny and fast-moving object detection in videos," master thesis, McGill Univ., 2021.
[19] K. Vats, M. Fani, D. A. Clausi, and J. Zelek, "Puck localization and multi-task event recognition in broadcast hockey videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4567–4575, 2021.
[20] H. Pidaparthy and J. Elder, "Keep your eye on the puck: Automatic hockey videography," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1636–1644, 2019.
[21] M. Yakut and N. Kehtarnavaz, "Ice-hockey puck detection and tracking for video highlighting," *Signal, Image Video Process.*, vol. 10, pp. 527–533, 2016.
[22] M. Li, H. Hu, and H. Yan, "Ice hockey puck tracking through broadcast video," *Neurocomputing*, p. 126484, 2023.
[23] K. Vats, W. McNally, C. Dulhanty, Z. Q. Lin, D. A. Clausi, and J. Zelek, "PuckNet: estimating hockey puck location from broadcast video," 2021.
[24] R. Cavallaro, "The foxtrax hockey puck tracking system," *IEEE Comput. Graph. Appl.*, vol. 17, no. 2, pp. 6–12, 1997.
[25] X. Duan, *Automatic determination of puck possession and location in broadcast hockey video*. PhD thesis, University of british columbia, 2011.
[26] J.-P. Parto, "Ice hockey checking detection from indoor localization data," Master's thesis, Tampere Univeristy, Finland, 2021.
[27] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8." https://github.com/ultralytics/ultralytics, 2023.
[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv Preprint arXiv:2207.02696*, 2022.
[29] Magnifi, "Automated sports highlights, AI video highlights." https://magnifi.ai/, 11 2023.
[30] Pendular, "Content generation technology for broadcasters and publishers." https://pendular.io/, 11 2023.
[31] Backlight, "Cloud-based video CMS, playout, distribution and app builder." https://www.backlight.co/streaming, 2023.
[32] WSC sport, "WSC sport." https://wsc-sports.com/, 2023.
[33] G. Gomez, P. Herrera López, D. Link, and B. Eskofier, "Tracking of ball and players in beach volleyball videos," *PLOS ONE*, 2014.
[34] Y. Liu, Q. Huang, Q. Ye, and W. Gao, "A new method to calculate the camera focusing area and player position on playfield in soccer video," in *Proc. Visual Commun. Image Process.*, International Society for Optics and Photonics, 2005.
[35] M. Majidi, M. H. Sarkhoosh, C. Midoglu, S. S. Sabet, T. Kupka, D. Johansen, and P. Halvorsen, "SmartCrop-H: AI-Based cropping of ice hockey videos," in *Proc. 15th ACM Multimedia Syst. Conf. (MMSYS)*, 2024.
[36] Swedish Hockey League, "SHL." https://www.shl.se, 2023.
[37] Y. Zheng and H. Zhang, "Video analysis in sports by lightweight object detection network under the background of sports industry development," *Comput. Intell. Neurosci.*, 2022.
[38] T. Soucek and J. Lokoc, "TransNet V2: an effective deep network architecture for fast shot transition detection," *CoRR*, 2020.
[39] B. Castellano, "Scenedetect." https://github.com/Breakthrough/PySceneDetect/tree/main, 2023.
[40] Z. Shou, J. Pan, J. Chan, K. Miyazawa, H. Mansour, A. Vetro, X. Giro-i Nieto, and S.-F. Chang, "Online detection of action start in untrimmed streaming videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 534–551, 2018.
[41] S. Tang, L. Feng, Z. Kuang, Y. Chen, and W. Zhang, "Fast video shot transition localization with deep structured models," *CoRR*, 2018.
[42] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process. (ICIP)*, 2002.
[43] J. Tang, X. Peng, X. Chen, and B. Luo, "An improved mobilenet-SSD approach for face detection," *Proc. 40th Chinese Control Conf. (CCC)*, pp. 8072–8076, 2021.
[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
[45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 779–788, 2016.
[46] S. Karakuş, M. Kaya, and S. A. Tuncer, "Real-time detection and identification of suspects in forensic imagery using advanced YOLOv8 object recognition models.," *Traitement Signal*, vol. 40, no. 5, 2023.
[47] M. H. Sarkhoosh, D. Sayed Mohammad Majidi, C. Midoglu, S. S. Sabet, T. Kupka, M. A. Riegler, D. Johansen, and P. Halvorsen, "SportsVision-YOLO: fine-tuned YOLOv8 model for soccer and ice hockey." https://github.com/forzasys-students/SportsVision-YOLO.
[48] LabelBox, "Data-centric AI platform for building and using AI." https://labelbox.com/, 2024.
[49] S. Saleem, M. Aslam, and M. R. Shaukat, "A review and empirical comparison of univariate outlier detection methods," *Pakistan J. Stat.*, vol. 37, no. 4, 2021.

[50] V. Crnojević, B. Antić, and D. Ćulibrk, "Optimal wavelet differencing method for robust motion detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2009.
[51] Węglarczyk, Stanisław, "Kernel density estimation and its application," *ITM Web Conf.*, vol. 23, 2018.
[52] S. S. Sabet, C. Midoglu, S. Z. Hassan, P. Salehi, G. A. Baugerud, C. Griwodz, M. Johnson, M. A. Riegler, and P. Halvorsen, "Comparison of crowdsourced and remote subjective user studies: A case study of investigative child interviews," in *Proc. Int. Conf. Quality Multimedia Experience (QoMEX)*, pp. 1–6, 2022.
[53] C. Midoglu, M. Hammou, A. Sharifi, L. Xing, M. Hasan, A. Storas, S. S. Sabet, S. A. Hicks, I. Strumke, M. A. Riegler, C. Griwodz, and P. Halvorsen, "Experiences and lessons learned from a crowdsourced-remote hybrid user survey framework for multimedia evaluation," *Encyclopedia Semantic Comput. Robot. Intell.*, vol. 0, 2023.
[54] S. Kemp, "Digital 2020: global digital overview." https://datareportal.com/reports/digital-2020-global-digital-overview, 2020.
[55] TelecOmmunication Standardization Sector of ITU, "P.910 - subjective video quality assessment methods for multimedia applications." https://www.itu.int/rec/T-REC-P.910.
[56] L. Sthle and S. Wold, "Analysis of variance (ANOVA)," *Chemom. Intell. Lab. Syst.*, vol. 6, no. 4, pp. 259–272, 1989.
[57] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.
[58] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *CoRR*, 2021.

MEHDI HOUSHMAND SARKHOOSH earned a Master's degree in Applied Computer and Information Technology from Oslo Metropolitan University, where he focused on applying artificial intelligence and machine learning techniques in sports analytics. His academic background includes Mathematics and mathematical modeling, which supports his work in developing AI-based solutions. He is currently an AI and ML developer at Forzasys AS, contributing to the implementation and development of machine learning models and AI applications. More information can be found at https://www.mehdi.cam.

SAYED MOHAMMAD MAJIDI DORCHEH is currently pursuing a Master's degree in Applied Computer and Information Technology at Oslo Metropolitan University. With a foundation in mathematical modeling and optimization, he is focusing on research related to the application of AI and machine learning in the sports industry.

CISE MIDOGLU is a postdoctoral fellow at the Simula Metropolitan Center for Digital Engineering (SimulaMet) and a developer/researcher at Forzasys AS. She received a B.Sc. degree in Electrical and Electronics Engineering from Bilkent University, M.Sc. degree in Information Technology from the University of Stuttgart, and PhD degree from the University of Oslo. She joined Simula Research Laboratory in 2016 and is currently working in the Department of Holistic Systems (HOST). Her research interests include AI/ML applications for sports multimedia systems, multimodal soccer analytics, streaming QoE, performance measurements in mobile broadband networks, and crowdsourcing. More information can be found at https://www.simula.no/people/cise.

SAEED SHAFIEE SABET is a developer and machine learning engineer at Forzasys AS. He earned his PhD from TU Berlin, where he delved deeply into his research interests, which include multimedia quality of experience, gaming, and virtual/augmented reality technologies. His work focuses on improving the user overall quality of experience in multimedia applications.

TOMAS KUPKA is COO at Forzasys AS. He has a PhD from University of Oslo in computer science looking at adaptive video streaming and performance. He is focusing on video streaming systems as a whole, including auto-generation of video highlights and summaries using AI.

DAG JOHANSEN is professor in computer science at UiT The Arctic University of Norway. His research interest is distributed computer systems, currently focusing on systems support for machine learning. He is exploring interdisciplinary research problems at the intersection of sport science, medicine, and law. A use-case receiving special attention is elite soccer performance development and quantification technologies as basis for evidence-based decisions. Focus is on intervention technologies where compliance and security are first-order concerns and design principles. Johansen has co-founded a series of startups, and he is an elected member of the Norwegian Academy of Technological Sciences.

MICHAEL A. RIEGLER is a chief researcher at SimulaMet and Professor at Oslo Metropolitan University. He is focusing on machine learning research applied to different application areas such as medicine and social sciences.

PÅL HALVORSEN is a chief researcher at SimulaMet, a professor at Oslo Metropolitan University, and a professor II at University of Oslo, all in Norway. At SimulaMet, he is the head of the Holistic Systems research department, which investigates challenges of complete end-to-end pipelines with particular focus on sport and medical applications. In this respect, his interests span areas in distributed systems and content analysis from both performance, efficiency and accuracy point of views. More information about authored papers, supervised students, teaching and community services can be found at http://home.simula.no/~paalh.