Instance Segmentation-Based Markerless Tracking of Fencing Sword Tips

Takehiro Sawahata[1] Alessandro Moro[2] Sarthak Pathak[3] and Kazunori Umeda[3]

***Abstract*— This study addresses the challenge of detecting the tip of a fencing sword. The swift motion and diminutive size of the fencing sword tip not only poses difficulties in detection but also occasionally lead to its omission from video recordings. Moreover, conventional detection approaches such as affixing markers to the sword tip are unsuitable in sports contexts as they could encumber the athletes. In light of these considerations, our research has devised a system that exclusively employs monocular camera images to consistently gather information about the sword tip. Even in cases where the tip is not captured, we propose a method for predicting its position based on historical data and subsequent interpolation. Specifically, the entire sword is recognized using instance segmentation. And the tip of the sword is identified with skeletal point information. In instances where the tip eludes detection, its position is projected using preceding information and skeletal wrist point data, to ensure uninterrupted tracking.**

Our proposed method's efficacy was confirmed through various experiments conducted under conditions mirroring actual match scenarios. These experiments demonstrate the effectiveness of our approach.

I. INTRODUCTION

In recent years, with the advancement of image processing technology, research in the field of fencing has been actively conducted. The visualization of game situations has been conducted to increase the attractiveness of the sport and offer more in-depth play analysis. Specifically, there are studies that identify footwork using skeletal point information extracted from images [1], as well as research that recognizes fencing swords in 3D and displays them in real-time on the wearer's AR goggles to support training [2], [3]. In "Sport: Fencing Matches AI [4]", Alexander P. attenpted an analysis of fencing match situations and identified the lack of information about the sword as a significant issue for making conclusive judgments. To address this challenge, J. Mo attempted to discern sword contact through an analysis of audio information from the match, acknowledging the difficulty of detecting swords via cameras [5]. However, it can be considered that relying solely on audio information is insufficient, as the game situation may change even sword contact or simply based on the sword's posture. From this, it is concluded that the acquisition and analysis of information about the sword are indispensable for the advancement of future research. Against this backdrop, this study focuses on the detection of sword tips. Fencing swords are very thin and the tip moves at high speeds. In past studies, detection using markers was common, but due to the burden on players during matches, actual use in competition was challenging. Therefore, efforts such as the development of Sword tracer [6], [7] by Takahashi *et al*, which involves attaching reflective material tape to the sword tip and tracking it with an infrared camera, have been attempted. However, this technology is not applicable to disciplines that do not require insulating tape. Alternatively, there is a system called "Fencing Tracking and Visualization System [8]", but it requires 24 4K cameras at 60fps, making it a massive setup and challenging to handle. Therefore, this research aims to develop a markerless tracking system without burdening players, using only a standard camera. These cameras face challenges such as the sword tip moving too quickly to be captured by the camera or the significant blurring of the sword tip as shown in Fig. 2b. To address these challenges, our method predicts and interpolates undetected parts using past sword information.

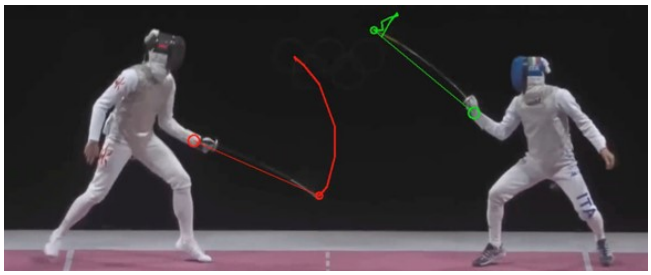
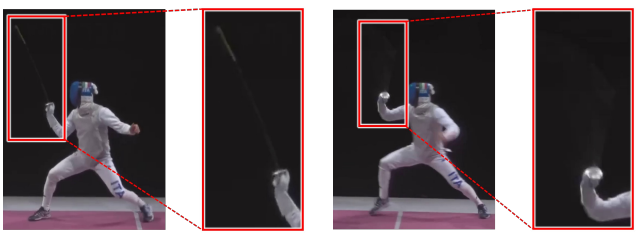


Fig. 1. Sword Tip Tracking: Visualization using our system.

The main contributions:

- Development of Markerless Tracking System using normal camera.



(a) Clear image (b) Blurred image

Fig. 2. Capturing the Tip of a Fencing Sword

[1]Precision Engineering Course, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan. sawahata@sensor.mech.chuo-u.ac.jp

[2]RITECS Inc., 3-5-11 Shibasaki, Tachikawa-shi, Tokyo, Japan.

[3]Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan. {pathak, umeda}@mech.chuo-u.ac.jp

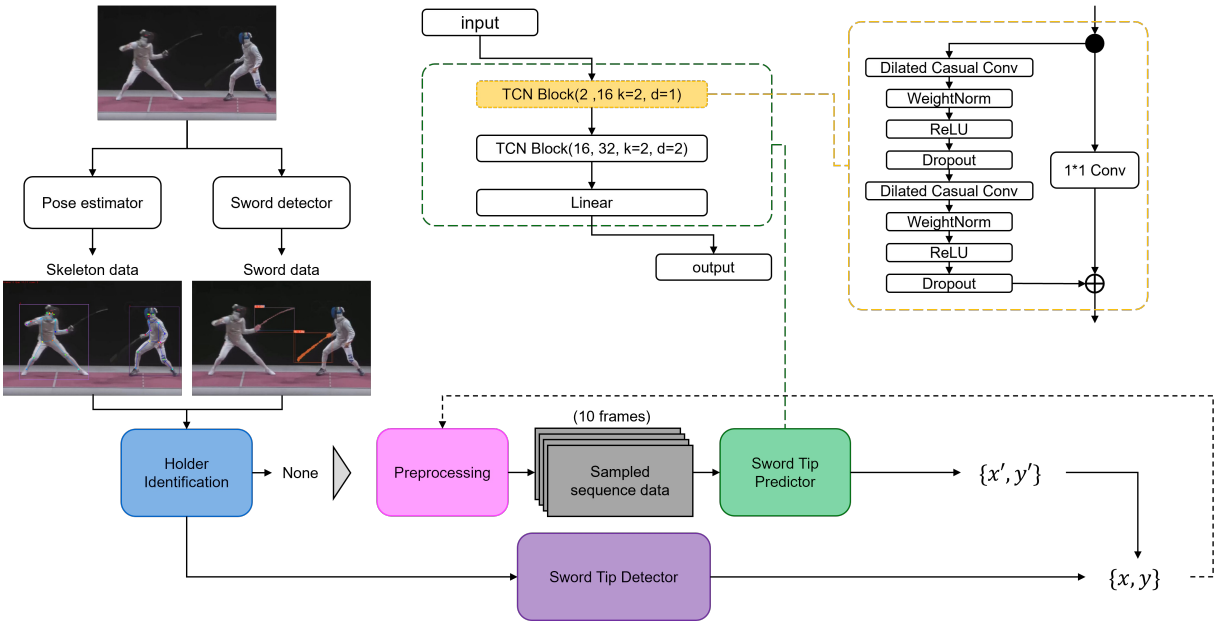


Fig. 3. Flowchart of proposed method.

II. METHOD

A. Concept

In this research, we utilize only RGB images to construct a detection system for the tip of a fencing sword. Given that it is impractical to place any burden on players during actual matches, we have adopted a markerless detection approach using only video footage. The system is designed to identify the position of the sword tip and to discern whether the sword is held by the player on the right or left side. Furthermore, the system addresses the challenge posed by the high-speed movement of the fencing sword tip, which can result in it not being captured in the footage. To overcome this issue, our system leverages past data and predictive algorithms to continuously obtain information about the sword.

B. Overview

In the proposed method of this research, four key processes are employed to achieve continuous detection of the fencing sword tip. As illustrated in Fig. 3, the first step involves detecting the entire sword from the whole image and obtaining mask information. Subsequently, to identify the holder, a correlation is made between the skeletal point information extracted from the image and the detected mask information. At this juncture, if the sword is undetected, the third step performs interpolation using a time-series neural network trained on the relative movement of the wrist. Conversely, when the sword is held, the final step identifies the sword tip by calculating the point farthest from the wrist coordinates within the identified holder's mask information. Through these processes, continuous detection of the fencing sword tip from video images becomes feasible. The following section provides a more detailed explanation of these four steps.

C. Swords Estimation

The sword detection process involves inputting RGB images into a network trained on a custom dataset. This output, consisting of mask tensors corresponding to the number of detected objects, serves as the final output for this process. For this detection, an instance segmentation technique called YOLACT++ [9], [10] is employed. This model, a simple convolutional model, concurrently generates prototype masks and predicts per-instance mask coefficients, thereby achieving rapid processing. Since the sword is characterized by its extreme slenderness and susceptibility to significant influence from the background environment, robustness is required. Hence, a highly robust instance segmentation method is adopted. The dataset comprises a total of 300 samples, all of which were acquired within a consistent environmental setting. However, these samples are categorized based on three distinct distance conditions, with each condition contributing 100 samples to the overall dataset. Furthermore, the annotation range is not limited to the sword alone, the area extending to the wrist is incorporated into the learning data as part of the sword.

D. Pose Estimation

In the approach being utilized for acquiring skeletal keypoint information, a top-down method is employed. The process begins by extracting the bounding boxes of individuals within a video using YOLOv5 [11], a renowned object detection model. This initial step ensures that the system can identify and isolate the human figures present in the frames of the video. Once the bounding boxes containing the individuals are identified, the next step involves the extraction of skeletal keypoints for two players. This is a crucial aspect of the method as it provides detailed information about the posture and movement of the players. To achieve this,

the ViTPose [12] model is used to obtain the keypoints within the previously detected bounding boxes. ViTPose, a transformer-based keypoint detector, is instrumental in accurately identifying the specific locations of various joints and limbs within the bounding box. By focusing on the confined area defined by the bounding boxes, the model is able to analyze and detect the skeletal structure of the two players with higher precision.

E. Holder Identification

We utilize the mask information of the swords and the skeletal keypoint information obtained in steps C and D to identify the players holding the swords. First, we differentiate the players engaged in the game from other people captured in the frame. Each player exhibits a brief pause of a few frames just before play begins. During this moment, we identify the forward-most hand on the wrist side as the "listening" hand. We keep track of this information until a score is determined. We recognize the sword by associating the sword's mask information within a specific distance from the wrist. This distance was determined empirically and functions as a critical factor in recognizing the sword. By applying this process to both the left and right players, we identify and distinguish the swords held by each player. The distance parameter was fine-tuned through practical observation, and this method is designed to be clear and straightforward. By focusing on the spatial relationships between the sword and the player's hand, our approach effectively identifies the sword bearers in a way that is both logical and easily understandable.

F. Sword Tip Identification

In the current step, we leverage the mask information obtained from the previous step, which identified the player, to pinpoint the location of the sword tip. We recognize the pixel within the mask that is farthest from the wrist coordinates as the tip of the sword and output its coordinates. This process ensures that only the information pertaining to the sword tip is extracted from the mask data.

G. Preprocessing

Preprocessing of Training Data for Sword Tip Prediction In the task of predicting the sword tip's position, preprocessing plays a vital role. The preprocessing consists mainly of two steps: "Position Correction" and "Scaling." These corrections are applied to each frame, retaining the information from the preceding 10 frames within 11 time steps.

1) Position Correction: The sword's coordinates are represented relative to the wrist's coordinates. Specifically, the position correction is achieved by subtracting the wrist's coordinates from the sword tip's coordinates. Since the fencing video oscillates left and right to track the athlete, this correction ensures that predictions can be made independent of the screen's movement or the athlete's position within the frame.

2) Scaling: To facilitate precise predictions irrespective of the distance between the camera and the athlete, scaling is performed by leveraging a moment of stillness just before the commencement of the fencing play. Specifically, the position values, after correction, are divided by the sword's length at that instant. This normalization allows for the prediction of the sword tip's position, regardless of the distance between the camera and the athlete.

Given the position p of the sword tip on axis c at time step t, the post-processed position $\tilde{p}$ is defined as follows (Equation 1):

$$\tilde{p}_c^t = \frac{p_c^t - p_{W,c}^t}{p_c^0 - p_{W,c}^0} \tag{1}$$

Here, W denotes the skeletal point of the wrist holding the sword. t represents the current time step, and $c \in \{x, y\}$ refers to either the x or y coordinate. The x and y coordinates of the sword's tip are utilized as inputs to the predictive model. $\tilde{\mathbf{p}}^t = [\tilde{p}^{t-11}, \tilde{p}^{t-9}, \ldots, \tilde{p}^{t-2}, \tilde{p}^{t-1}]$

H. Sword Tip Predictor

The Sword Tip Predictor consists of 2 Temporal Convolutional Networks (TCN) [13] with dilated causal convolutions to handle the 2D coordinates of the tip of a fencing sword as a time series, predicting missing coordinate positions.

1) Input Layer: The model's input consists of 10 frames of 2D coordinates of the sword's tip. The shape is $[B, 2, 10]$.

2) Temporal Convolutional Layers: The Temporal Convolutional Layers employ dilated causal convolutions to capture long-range dependencies in the time series without violating the causal order of the data. They are composed of Temporal Blocks defined as follows:

$$h_1 = \text{ReLU}(\text{Conv1d}(x, W_1, d) + b_1) \tag{2}$$

$$h_2 = \text{ReLU}(\text{Conv1d}(h_1, W_2, d) + b_2) \tag{3}$$

$$r = \text{downsample}(x) \text{ if required} \tag{4}$$

$$y = \text{ReLU}(h_2 + r) \tag{5}$$

Here, x is the input, W_1, W_2, b_1, b_2 are the weights and biases for the convolutional layers, and d is the dilation factor. The downsample operation is performed if required.

3) Output Layer: The output layer predicts the next frame's x and y coordinates of the sword's tip. The shape of the output is $[B, 2, 1]$. The formula to calculate the coordinates from the last Temporal Block's output is as follows:

$$\text{output} = \text{Linear}(h_{\text{last}}, W_{\text{out}} + b_{\text{out}}) \tag{6}$$

Here, h_{last} is the output from the last Temporal Block, and W_{out} and b_{out} are the weights and biases for the linear layer.

III. EXPERIMENTS

A. Experiment on Sword Tip Predictor

In this study, we evaluated the performance of the Sword Tip Predictor (STP). We will refer to Sword Tip Prediction" as STP in this paper As a benchmark, we conducted comparative experiments with a linear regression model. Data collection was performed using the V120: Trio system to

capture 2D sword movements. Fig. 4a depicts a standard sword, Fig. 4b shows the swordtip equipped with a motion capture device, and illustrates Fig.4c the attachment process. A specific component (indicated as 3 in Fig. 4c) was manufactured using a 3D printer. This component includes an upper part with a screw-type design that enables secure attachment of the marker. This design ensures continuous and stable data collection even during rapid movements or twisting of the sword. For capturing wrist position data, we utilized fencing gloves. Taking advantage of a portion of the glove that is made of Velcro, we securely attached the marker base to the glove(Fig. 4d).

The collected 120[fps] data was downsampled to a frame rate of 30[fps] to synchronize with video recordings. The dataset was organized in sets of 10 frames each, resulting in a total of 3600 sets. The distance between the camera and the sword was set within 4.0-5.0[m], and data was collected in four configurations by changing the hand holding the sword and its orientation. The distribution of these four types of data is uniform. The STP was trained over 10,000[epochs].

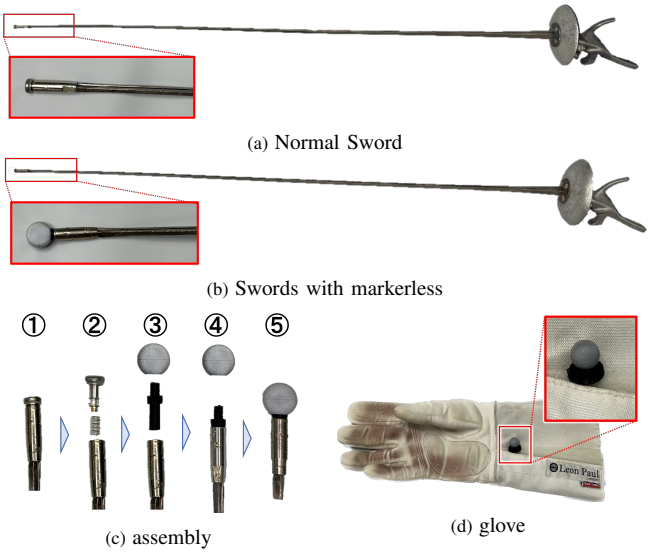
Table. I presents the experimental results using 100 data for each frame length. These 100 data consist of time-series data of the x and y coordinates of the sword's movement. Similar to the training data, the test dataset is composed of four configurations. The term "Frame Length" refers to the number of frames included in each data set for analysis. In your table, you have frame lengths of 3, 5, and 10. This means that each data set used for the TCN and Linear models consists of sequences of 3, 5, or 10 frames, respectively. The term "Accuracy" in the table refers to the percentage of correctly predicted sword tip positions within different pixel ranges (0-15[pix], 0-30[pix], and 0-40[pix]). Higher percentages indicate better model performance. For example, in the case of TCN with a frame length of 10, the accuracy reaches 100[%] for the 0-40[pix] range. Across all data lengths, the Temporal Convolutional Network (TCN) demonstrated higher accuracy compared to the linear regression model. Notably, the longest data length, TCN10, exhibited the highest accuracy rate. These results confirm that the method employing TCN can predict sword tip movements with high precision.

B. Experiment on whole system

In this study, experiments were conducted across the entire proposed system. The evaluation was carried out over a total of 5 scenes, with 60 frames for each scene. Each scene

TABLE I. Results of Sword Tip Predictor

Method	Frame Length	Accuracy[%]		
		0-15[pix]	0-30[pix]	0-40[pix]
TCN	3	89	94	95
Linear		72	87	87
TCN	5	91	94	100
Linear		72	81	87
TCN	10	92	96	100
Linear		69	74	86



(a) Normal Sword

(b) Swords with markerless

(c) assembly

(d) glove

Fig. 4. Capturing the Tip of a Fencing Sword

Fig. 5. Environment

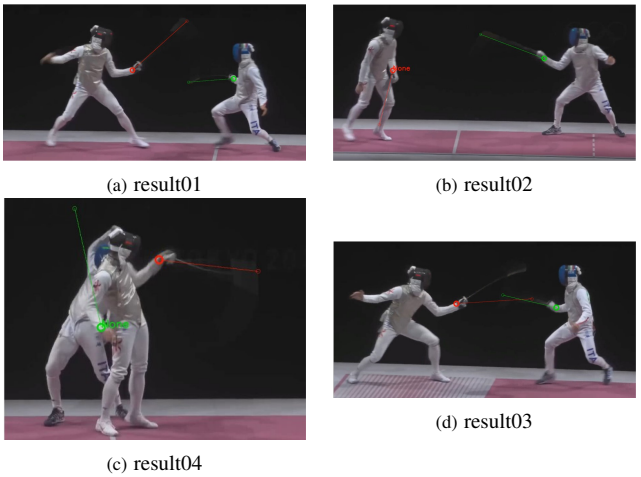
featured two athletes and was randomly extracted from actual game footage. The ground truth was determined by our research team through careful observation of the video. For frames where the tip of the sword was not visible, the next frame was consulted or the location was predicted based on empirical rules.

The results are summarized in Table. II. A 20[pixel] margin was considered as the criterion for accurate detection. This margin was chosen based on the observation that multiple individuals, when annotating the 60-frame videos, had an average discrepancy of 21 pixels when assigning ground truth data. Therefore, in this experiment, predictions within a 20-pixel margin from the ground truth were considered accurate. The average accuracy of our method was found to be 88.7%, indicating an improvement in accuracy compared to methods without the incorporation of STP. Furthermore, when compared to linear regression models, STP demonstrated an enhancement in accuracy across all evaluation metrics.

These findings confirm that the introduction of STP contributes to the improvement of position estimation accuracy, thereby validating the effectiveness of the proposed method in this study.

TABLE II. Results of Experiment on whole system

Scene	Player	Accuracy[%]
1	Right	100.0
	Left	100.0
2	Right	86.3
	Left	93.3
3	Right	80.0
	Left	81.7
4	Right	100.0
	Left	100.0
5	Right	70.0
	Left	75.0



(a) result01 (b) result02

(c) result04 (d) result03

Fig. 6. result

C. Discussion

1) Experiment on Sword Tip Predictor: Our method outperformed linear regression models across all metrics. Additionally, the results from the Temporal Convolutional Network (TCN) indicated stable performance regardless of the length of the input vector. We observed that the method was capable of accurately predicting sword movements in scenes with both intricate and large-scale motions, irrespective of data seasonality. On the other hand, linear regression models only provided high-accuracy predictions for trajectories with seasonal patterns that matched the length of the vector. Furthermore, as the length of the vector increased, the complexity of the sword movements also increased, leading to a decrease in prediction accuracy.

2) Experiment on whole system: In Experiment on whole system, we evaluated the proposed method using scenes extracted from actual game footage, achieving an average accuracy rate of 88%. The photos of the demonstration are presented in Fig. 6. Fig. 6a shows the output without predictions, while Fig. 6b represents the output with the applied prediction model. In cases where the prediction model is used, the wrist coordinates are indicated as "None."Unsuccessful scenes are illustrated in Figure 10. The errors could be broadly categorized into three types: "misidentification of skeletal points(Fig. 6c)," "significant overlap of swords (Fig. 6d)," and "incorrect prediction values." Errors in skeletal point identification fundamentally altered the reference wrist position, subsequently affecting the output information for the sword tip. In cases where the swords significantly overlapped, the algorithm sometimes incorrectly identified the opponent's sword hilt or wrist as the sword tip. These were considered false detections, attributed to the inability to output the correct detection. Lastly, the incorrect prediction values are believed to have been influenced by the inaccuracies in the sword detector. During the training phase of STP, only ground truth data were used without introducing any noise into the training process. As a result, there is a possibility that errors in prediction output increased due to the lack of variability in the training data.

IV. CONCLUSION

Our proposal presents a sword tip detection system for fencing competitions that exclusively employs image data. Specifically, we utilize real-time semantic segmentation to detect overall sword information and skeletal points, enabling the identification of each player's sword tip. Given the small size and high-speed movements of fencing sword tips, instances arise where they do not appear in the images. Even in such scenarios, our approach employs a sword tip predictor to predict and interpolate the undetected tips based on past detections. This methodology allows for the detection of sword tips that are not visible in the images. The effectiveness of our system was validated through evaluation experiments utilizing actual match videos. In the accuracy evaluation experiments of the Sword Tip Predictor (STP), our network, which incorporates Temporal Convolutional Networks (TCN), outperformed other methods, achieving the highest scores when predicting for ten frames. When utilizing STP within the comprehensive system experiment, we achieved an accuracy of 88.7%. However, challenges remain in our approach, particularly regarding accuracy drops in scenarios where fencers overlap or sword-to-sword contacts occur. To address these challenges, we are considering the development of a specialized skeletal point estimation system tailored to fencing competitions. Furthermore, enhancing the precision of STP using datasets that include instances of sword contact is essential.

Looking ahead, we envision the utilization of these detection results to develop an AI referee capable of assessing the state of fencing matches. Creating a rule-based AI referee requires accurate determination of sword-to-sword contacts, a critical aspect of judging. While our current system alone cannot discern such contacts, by combining continuous sword information and posture data extracted from images, we strive to achieve a rule-based AI referee capable of making these determinations.

REFERENCES

[1] K. Zhu, A. Wong, and J. McPhee, "FenceNet: Fine-grained Footwork Recognition in Fencing," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun. 2022, pp. 3588–3597.

[2] F. Malawski, "Real-Time First Person Perspective Tracking and Feedback System for Weapon Practice Support in Fencing," Jan. 2018.

[3] F. Malawski, "Immersive Feedback in Fencing Training using Mixed Reality," Computer Science, vol. 23, no. 1, Art. no. 1, Mar. 2022.

[4] Pageaud. A (2019, September 14), "Sport : Fencing Matches AI. Kaggle," https://www.kaggle.com/datasets/alexpgd/sport-fencing-matches-ai / Accessed July. 10, 2023.

[5] J. Mo, "Allez Go: Computer Vision and Audio Analysis for AI Fencing Referees," Journal of Student Research, vol. 11, Nov. 2022.

[6] M. Takahashi *et al*, "Sword tracer: visualization of sword trajectories in fencing," in ACM SIGGRAPH 2018 Talks, Vancouver British Columbia Canada: ACM, Aug. 2018, pp. 1–2.

[7] M. Takahashi *et al*, "Real-time visualization of sword trajectories in fencing matches," Multimed Tools Appl, vol. 79, no. 35, pp. 26411–26425, Sep. 2020.

[8] Y. Hanai *et al*, "Fencing tracking and visualization system," in SIGGRAPH Asia 2021 Real-Time Live!, in SA '21. New York, NY, USA: Association for Computing Machinery, Dec. 2021, p. 1.

[9] D. Bolya*et al*, "YOLACT: Real-time instance segmentation," Proc. ICCV, pp. 9156-9165,, Oct. 2019.

[10] D. Bolya*et al*, "YOLACT++ Better Real-Time Instance Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 1108-1121, Feb. 2022.

[11] G. Jocher *et al*, "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," Zenodo, Nov. 22, 2022.

[12] Y. Xu *et al*, "ViTPose: Simple vision transformer baselines for human pose estimation," arXiv:2204.12484, 2022.

[13] Colin Lea *et al*, "Temporal convolutional networks for action segmentation and detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.