

Video Based Fencing Analysis: A Pilot Study of Action Classification and Short-Horizon Forecasting

Avyang Sai Devineni¹

Received April 26, 2026

Accepted September 25, 2026

Electronic access October 31, 2026

Fencing movements happen very fast and depend on body dynamics, distance, and the scenario of the moment. In this study, the adequacy of using the pose trajectories generated by a single visible fencer in the analysis and recognition of the movement is examined. Four training clips from workbooks were used, providing the normalized coordinates like Media Pipe, as well as training, validation, and testing datasets for each clip. Four more full fights were used for validation and testing with frames being captured once every five frames. Two humans annotated the action segments, and a stable-tracking filter found the corresponding pose trajectory for each segment. Exploratory analysis of coordinate coordination through correlations, three one-frame right-wrist forecasting baselines, current-segment multilayer perceptron (MLP), and previous window LSTM was performed. For the complete-video test set, the MLP could classify 155 out of 227 segments correctly, obtaining 68.28% accuracy, 66.78% balanced accuracy, and 66.16% macro F1. With a candidate-rule strategy, the baseline accuracy was 45.81%. The minimal MAE value for horizontal forecasting of right-wrist position was 0.005991 normalized units or approximately 0.6% of the image width, mainly due to smooth transition between consecutive frames. Sensitivity study of the LSTM neural network with previous window sizes equal to 15, 30, 60, and 90 samples showed preference of 30 samples from the viewpoint of validation macro F1 score. The chosen LSTM had accuracy of 28.44%, balanced accuracy of 28.55%, and macro F1 of 23.36% for labels of next segments. Results allowed conducting pose-based analysis of currently performed fencing moves and sequence of immediate coordinates. In the future, tactical forecasting and next actions prediction will be improved by adding more bouts, tracking of two fencers, blade coordinate information, correct timing, and definition of independent segments.

Keywords: data science; fencing; pose estimation; action classification; coordinate forecasting; stable tracking

Introduction

Fencing is analyzed as a field in which fast movement is complemented by constant tactical decision-making. Pose estimation from videos is used to represent observable landmarks numerically for trajectory analysis frame-by-frame. The purpose of this work is to determine how well these trajectories are suitable for analysis in a small and reproducible fencing dataset.

The study includes three main analyses. Firstly, the pose trajectory analysis is aimed at finding the correlation between movements of selected coordinates in the shoulder, wrist, knee, and ankle. Secondly, short-horizon coordinate forecasting seeks to predict the horizontal position of the right wrist one frame into the future. Thirdly, the current-segment classifier uses pose-derived features for assigning one of four labels: en garde, lunge/attack, retreat, or transition/other to the segment. Furthermore, the pilot LSTM is conducted to compare several prior-window lengths and use 30 samples of poses selected in the validation process for classification of the up-

coming segment.

The data contains clips from the original workbook, as well as four full sessions for verification and testing on a video level. In each case, only one split of the video is taken. A few of the questions that arise during this study include whether it is possible to generate simple coordinate patterns; if there can be any predictions of wrist movements; and if pose summarization is enough to classify the actions within this particular session. The findings of this research are referred to as pilot findings.

Related Work

The methodological background for this study is markerless pose-based motion analysis. Specifically, Bazarevsky et al.¹ present BlazePose that generates the 33 keypoints that define the MediaPipe pose used in this study². Chida et al.³ compared the results of 2D video-based pose analysis and 3D motion capture regarding hip, knee, and ankle angles during fencing lunges and found out that the highest agreement corresponded to sagittal knee angles, thus explicitly supporting the

¹ Cambridge Center for International Research

Table 1 Original clip provenance and metadata.

Field	Clip 1	Clip 2	Clip 3	Clip 4
Clip name	csv_data_cannone4	csv_data_sikloski4	csv_data_cannone3	csv_data_sikloski3
Frames	1,179	1,153	1,392	1,396
Split	Train	Train	Validation	Test
Source type	Broadcast pose sequence	Broadcast pose sequence	Broadcast pose sequence	Broadcast pose sequence
Four-class human-reviewed action segments	59 segments	35 segments	80 segments	76 segments
Weapon	Epee	Epee	Epee	Epee
Athlete handedness	Right	Right	Right	Right

application of 2D video-derived coordinates in fencing. Furthermore, markerless pose-estimation algorithms have been validated against marker-based motion tracking in the context of other sports^{4–6}, specifically, in an automated multi-view workflow⁷, single-camera pose estimator validated against VICON-ground truth⁸, and automatic 2D keypoint and event detection in sports video⁹. All these studies support error, agreement, and reliability validation within their tested methods and tasks; however, they do not validate either MediaPipe normalized screen coordinates or Pearson correlations as a diagnostic method for fencing actions. In this paper, MAE and RMSE measure the forecasting error in the coordinates, while Pearson correlations retain descriptive function only.

Fencing competition video and fencing biomechanics literature provide the background to the coordinates analyzed in this pilot. Specifically, Magnani and Defrasne Ait-Said¹⁰ reconstruct footwork, mutual distance, and lunge motion from Olympic competition video. Guan et al.¹¹ analyze the relation of the lunge speed to lower extremity forces and joint mechanics, while Bottoms et al.¹² relate lower extremity lunge kinematics to weapon velocity in experienced epee fencers. Akbas et al.^{13,14} describe early and anticipatory postural adjustments preceding the lunge, and Borysiuk et al.¹⁵ analyze muscle activation, ground reaction forces, reaction time, and movement time during the preparatory and execution phases of the lunge. Thus, all these studies emphasize the importance of observable wrist and lower limb trajectories and temporal preparation to an action; however, they do not validate any correlation thresholds, labels, or chosen LSTM window applied in this paper.

In addition, three papers specifically related to the temporal and classification task analyzed in this pilot can be distinguished. First, Honda et al.^{16,17} use recurrent networks that model the interaction of two players to predict the responses to the opponents' thrust (including retreating). Second, Malawski and Kwolek¹⁸ classify six pre-segmented footwork actions using depth, skeleton, and inertial cues, while Malawski and Krupa perform per-frame classification and temporal segmentation of four footwork actions using RGB pose estimation¹⁹. Third, Zhu et al.²⁰ propose FenceNet –

a temporal-convolutional network for fine-grained footwork recognition using 2D pose data on 652 videos and Lai et al.²¹ suggest a transformer-based approach for fine-grained fencing and boxing actions classification based on raw videos²². These studies establish precedents for the interaction-aware prediction, footwork recognition²³, temporal segmentation²⁴, and pose extraction in fencing²⁵; however, the current pilot focuses on coordinate and segment labeling tasks.

Fencing Background

Fencing is an Olympic combat sport played between two contestants on a narrow track known as piste. There are four stages at which the touch is made including preparation, attack, defense, and recovery. The foil, epee, and sabre sports use different target zones as well as scoring techniques.

As a sport that can be described as physical chess, fencing is a game in which every move restricts the opponent's possible actions. Strategic moves arise in the dynamic setting of the sport through movement of athletes who change distances²⁶, timings, blade position, and their footwork among other elements. In the present study, the recorded data produce only one body pose sequence that corresponds to one of the athletes.

Methods

Dataset

Initially, there were four pose sequences from one fencer obtained from Fencing Data Spreadsheet.xlsx that was taken from a set of three YouTube videos^{27–29}. In total, this provides 5,120 indexed frame samples, where each sample consists of the coordinates (x, y, z) and visibility for 33 landmarks. The video durations, data splitting, and metadata that can be recovered are provided in Table 1.

Since there is only one easily identifiable fencer in each video clip, there is less ambiguity regarding single person pose extraction in both training and test cases. The poses are analyzed using automatic landmarks detected

Table 2 Complete-video allocation for validation and testing.

Split	Video ID	Source video	Holdout role
Train	MVqBp6dDTXg, Dn8-M2zW9PU	ICP-lrvsR44, Existing paper source-video pool ^{27–29}	Training/development source pool already used for the current pilot
Validation	aVtn1 19IU4	Budapest World Cup 2026 MS - L64 - Aron Szilagyi HUN v Cosimo Bertini ITA ³⁰	Complete held-out validation bout
	UO9OaqHcAIU	Budapest World Cup 2026 SMS - L16 - Oh Sanguk KOR v Remi Garrigue FRA ³¹	
Test	MMN1tvPzNzM	Turin Grand Prix 2026 SMF - GOLD - Nick Itkin USA vs Guillaume Bianchi ITA ³²	Complete held-out test bout
	rD9rm-MmAE4	CIP World Cup 2026 SMF - GOLD - Ryan Choi HKG v Alex Massialas USA ³³	

and saved in a workbook. The coordinates are normalized according to the scale of the workbook and follow the frame order. The labels of workbook sheets are always csv_data_cannone4, csv_data_sikloski4, csv_data_cannone3, and csv_data_sikloski3. csv_data_cannone4 and csv_data_sikloski4 form the training split, csv_data_cannone3 forms the validation split, and csv_data_sikloski3 forms the test split. Since the clips are separated, there is no temporal leakage in training. However, since there are no original video windows and crop values available, the cross-frame consistency of athlete choice cannot be checked independently.

Complete-Video Holdout Expansion Protocol

Since each of the original validation set and test set had one processed video, the action-classification task involved four complete video bouts: two validation video bouts and two final testing video bouts. All frames, segments, annotations, and model samples from one source video were placed into one split.

This complete-video holdout strategy involves two video bouts in validation split and two video bouts in the final test split, which means that all the segments of one video will be in the same split. It leaves less data for development but makes sure that the final evaluation is based on two separate video bouts. The alternative of 3/1 would leave more data for development but make the testing results depend on the athletes, conditions of broadcasting, and the distribution of classes in one bout.

Complete video broadcasts include not only close-ups, but also two fencers along with officials, replays and pauses. Thus, the one-person detection system can switch among visible people. As a result, the described process of stable track generation gives one continuous sequence of poses for each annotation segment. This method acts as identity continuity filter instead of named athlete or two-person tracking, the latter being reserved for the future work.

Four bouts gave 385 candidate segments for inspection by

two annotators and assisted by rules. In current-segment model 250 segments from original clips were used for training, 157 segments from complete videos for validation, and 227 segments from complete videos for testing.

Ethical Use of Broadcast Footage

The dataset involves freely available fencing streams on YouTube, which are cited in the Dataset section. The pose coordinates and actions labels obtained from the videos were analyzed instead of using privately obtained videos or the names of athletes. Redistribution of the original video streams is prohibited. Further use is expected to be consistent with the platform rules, privacy of the athletes, and any other requirements of the rights holders.

Human-Verified Action Labels

The action segments were generated from the pose coordinate information based on deterministic principles involving center-based motion, right wrist extension, right wrist velocity, stance width, and short run smoothing. The segment boundaries were marked at those points where the candidate label would change after the smoothing process to result in 254 action segments in the four videos. For the analysis of current segments, there will be 250 segments in four classes. Two different human annotators, annotated 001 and 002, evaluated the segment windows of the same video against the context of the source video and labeled the segments using labels from the action taxonomy table.

Table 3 Four-class action-label definitions.

Label	Operational definitions for the analysis
en_garde	Stable and balanced guard stance with no or minimal forward and backward displacement
lunge_attack	Forward movement and attack with body displacement toward the opponent along with arm extension or expansion of the base
retreat	Backward displacement from the opponent after preparation for the attack or the actual attack and exchange
transition_other	Ambiguous, transitional, recovery, or background movement outside the three action-specific labels

The proposed candidate rules use per-clip quantiles, which allow the threshold values to vary depending on the camera

resolution, the athlete's position, and orientation of the video. Every candidate rule evaluates the motion in terms of five-frame windows. No motion along with stable posture implies en garde; wrist extension along with forward motion implies lunge or an attack; backward movement of the body implies retreat. All other kinds of motion belong to the category of transition/other. A short filtering phase is used to filter out spurious labeling instances. A total of 1,621 frames were evaluated with 885 frames being relabeled.

Details of Pose Extraction and Preprocessing

Holdouts of complete videos were taken with OpenCV from source videos with frame rates of 50, 50, 53, and 50 fps. Strides of 10, 10, 11, and 10 fps resulted in sampling rates of 5.000, 5.000, 4.818, and 5.000 fps, respectively. Pose was extracted with MediaPipe 0.10.14 through single-person Solutions Pose API using no static image mode, model complexity set to 1, landmark smoothing, no segmentation, and detection and tracking confidence threshold of 0.5. This API returns only one pose per frame, and num_poses parameter is absent. The complete RGB frame was utilized without any cropping or resizing. The landmarks of the image were normalized such that x, y are measured relative to the width and height of the frame while z is the relative depth with no clip-wise normalization.

Frames with no detected poses were kept with empty landmark fields and skipped from pose features analysis. Frames with detected poses contained all 33 x, y, z, and visibility values of landmarks. Stable-tracks filter classified a sample as unstable if the mean visibility of the right shoulder, wrist, hip, knee, and ankle was less than 0.25. The filter split a track if horizontal or vertical displacement of body center was higher than the maximum of 0.18 normalized units or median absolute displacement plus 6 times median absolute deviation of the track. In each annotated subsegment, the classifier made use of the longest continuous sequence possible. No further correction for occlusion, motion blur, cuts, replays, or changes in the camera angle was performed; nonetheless, missed detections, visibility filtering, and splitting of tracks served as indirect protection against these issues. The workbook used in the study contains normalized landmark values.

Correlation Analysis

Pearson correlations were used as descriptive statistics for describing the covariation pattern between chosen normalized pose coordinates per each selected video clip. Since the frames were not independent from each other, the autocorrelations of lag-1 for both sequences were used to replace the number of frames with the effective number of samples per each correlation that should be considered. The Fisher Z trans-

formation resulted in 95% confidence intervals and two-sided p-values. The Benjamini-Hochberg method was applied to adjust for the false discovery rate within the 12 planned comparisons. The magnitude of the correlation was classified as low ($|r| < 0.30$), medium ($0.30 - 0.49$), or high (≥ 0.50). Such classification reflects the size of an effect but not the statistical significance which was tested by excluding zero from the confidence interval and p-value less than 0.05.

Evaluation

The coordination forecasting method uses a forecasting horizon of $h = 1$ frame and evaluates the performance of the three estimates of the right-wrist x-coordinate: training mean, persistence, and constant velocity. The error measures MAE(h) and RMSE(h) are computed using N forecast pairs per split; N is provided in Table 5. Current action classification is analyzed by accuracy, balanced accuracy, macro F1, weighted F1, and confusion matrix.

The candidate labels are evaluated against each annotator's label in shared segment windows; for annotator 002, the evaluation against annotator 001 is done using percent agreement and Cohen's kappa measure. The MLP used in current segment was selected based on complete video validation bouts, while the complete video test bouts were used for evaluation of result.

Mathematical Formulation

Each frame contains 33 pose landmarks. The frame prediction task focuses on the right wrist, which corresponds to landmark 16. Its x-coordinate is given as a normalized value, where the difference of 0.01 corresponds to 1% of the image width. Let $x(t)$ denote the ground truth wrist location at frame t , and $\hat{x}(t+h|t)$ denote the prediction at frame t for frame $t+h$. The forecast horizon h is measured in frames and is equal to 1 in our experiment, meaning that every sequence gives a prediction for the following frame. There are three simple baselines showing the predictability of the following frame. The first baseline is training mean. It uses the mean of the right wrist location in the training sequences. The second baseline is persistence. It predicts that the following position will be the same as the current one:

$$\hat{x}(t+1|t) = x(t)$$

The constant-velocity baseline continues the most recent frame-to-frame change:

$$\hat{x}(t+1|t) = 2x(t) - x(t-1)$$

For forecast i , made at frame t_i , define the error as $e_i(h) = x(t_i+h) - \hat{x}(t_i+h|t_i)$.

$$MAE(h) = (1/N) \sum |e_i(h)|$$

$$RMSE(h) = \sqrt{[(1/N) \sum e_i(h^2)]}$$

In both expressions, the sum is calculated for $i = 1$ to N . In this case, h is the forecast horizon, N is the number of possible forecasts pairs in the split, and t_i is the frame where forecast i starts. Both functions are calculated in normalized image-coordinates, where lower values indicate better estimates. For the constant-velocity case, we have slightly less usable data, since it requires also knowing $x(t_i - 1)$.

These two classifiers answer different questions and have different time intervals. The multilayer perceptron (MLP) gets the summary features of frames included in some action segment and classifies it to the action currently performed. On the other hand, the long short-term memory network (LSTM) gets the pose samples before the start of the following action segment and classifies it to the action that is supposed to be performed there. We have tested several window sizes (15, 30, 60, 90 samples) and settled for 30 samples using the validation macro F1 score. The window size is the input context, and LSTM provides only one future action label. Concisely, the MLP recognizes an action that has started, while LSTM tries to predict the next action.

Short-Horizon Coordinate Forecasting

The present research project employs a single-frame increment ($h = 1$ frame), which makes the scope of the issue under investigation narrow, what is the degree of approximation of the wrist coordinate by an elementary rule in the following frame? The source videos are played at the speed of 25-30 frames per second (each frame equals 33-40 ms). The results obtained in the workbook are measured in frames. It would be easy to create the correlation between each worksheet and its source video window and measure everything in milliseconds. Therefore, the current experiment is a good example of the immediate coordinate continuity assessment³⁴.

Action Classification for Stable Track Current-Segment Action Classification

The latest action classification experiment is reported using a reproducible current-segment model with four-class verified action labels. Each example consists of pose-derived frames limited to the action segment, rather than frames prior to the next segment. Paper-clip segments result in 250 training examples, where the possible-parry rows are removed. In the complete-video holdout experiment, there are 157 validation examples obtained from two complete validation bouts used for model selection and 227 test examples from two complete test bouts.

Each action segment is described with a feature vector. 12 pose features include mean, standard deviation, min, max,

first, last, change, and range of the feature. The pose features consist of body position, wrist position and angle, stance width, knee-ankle distance, recent displacement, wrist speed, stance change, movement amount, and visibility. Also, the vector contains the number of frames, number of frames in stable-tracks, confidence in the candidate, and candidate label. Complete-video segments are based on the longest continuous pose track meeting the stability criteria. All features are normalized only based on training data.

The classification model consists of one hidden layer with a *tanh* activation having 64 neurons and four-class softmax output. Training procedure uses weighted cross-entropy loss with class weights, L2 regularization with coefficient of 10^{-4} , batch size of 32, learning rate of 0.01, gradient clipping with the norm of 5.0, and fixed random seed of 42. The early stopping criterion ends the training after 120 epochs without improvement in validation macro F1. The validation model is trained using the original clips and validated on the complete-video validation bouts. Final model is trained on the expanded dataset by including validation labels and validated on the test bouts once.

Prior-Window LSTM with Four Classes Context

The Long Short-Term Memory (LSTM) provides the contextual information for the original temporal query. The selected architecture uses 30 pose samples from before the action segment, where the target is the verified label of the segment. The four classes are: en garde, lunge/attack, retreat, and transition/other. It differs from the MLP, as the latter uses the summary of poses inside the current segment. Therefore, Table 7 shows two separate tasks.

The LSTM uses the same source-video partitioning as in Table 2. Workbook clips with four actions gave 246 training windows. Video segments aVtn119IIU4 and UO9OaqHcAIU gave 155 validation windows for model selection. Video segments MMN1tvPzNzM and rD9rm.MmAE4 provided 225 windows for final testing. The first segment of each video was discarded due to the absence of the previous context window, and parries targets were excluded from the four-class experiment.

The LSTM consists of one recurrent layer with a hidden size of 16. Linear softmax head maps the last hidden vector to the four class probabilities. Each frame contains 12 pose-based features. Training was done using Adam optimizer with a learning rate of 0.005, batch size of 16, 200 epochs as the maximum number, class-weighted cross-entropy loss, gradient clipping with the norm of 5.0, and fixed random seed of 42. The early stopping stopped training after 30 epochs without validation macro F1 improvement. Dropout was set to 0.

Sensitivity analysis of the input window with 15, 30, 60, and 90 pose samples was performed under the same four-class

scheme. The validation macro F1 scores are 22.63%, 24.97%, 24.52%, and 24.86% respectively, leading to the choice of the 30-samples window for further test evaluation. In complete video with 5 Hz sampling rate, the windows contain approximately 3, 6, 12, and 18 seconds of data respectively. The chosen 30 samples window is the empirical pilot test value window and is not optimized because of physiological factors. The precise time of workbook window cannot be established since there are no source-video windows associated with workbook sheets.

Temporal Dependence

Consecutive video frames tend to have strong similarities because of the smoothness of the body's motion. Therefore, prediction one frame ahead becomes an easier problem compared to multiple frames and/or future actions. Clip-level data division makes sure there is no confusion in distinguishing bouts between different datasets, but the proximity of consecutive frames is an inherent aspect of the problem. Therefore, Table 5 should be treated as a measure of continuity. For long-term forecasts, validated frame rates, time-specific horizons, and models with more movement history can be beneficial.

Results

The correlation coefficient for vertical movements of shoulder-wrist in the four clips varied from 0.4419 to 0.8752, for horizontal movements of knee-ankle from 0.9560 to 0.9656, and for horizontal-vertical wrist movements from -0.2892 to 0.0421 . Effect sizes and inferences adjusted by the autocorrelation of each pair of clips are listed in Table 4.

Table 4 Autocorrelation-adjusted inference for selected coordinate correlations.

Clip	Coordinate pair	r	Effective n	95% CI	BH-adjusted p
csv_data_cannone4	wrist x-y	-0.2877	28.5	[-0.5944, 0.0920]	0.162
csv_data_cannone4	shoulder-wrist y	0.8010	32.9	[0.6310, 0.8976]	<0.001
csv_data_cannone4	knee-ankle x	0.9656	32.2	[0.9303, 0.9832]	<0.001
csv_data_sikloski4	wrist x-y	0.0421	107.5	[-0.1485, 0.2297]	0.667
csv_data_sikloski4	shoulder-wrist y	0.8752	158.1	[0.8328, 0.9073]	<0.001
csv_data_sikloski4	knee-ankle x	0.9560	22.6	[0.8967, 0.9816]	<0.001
csv_data_cannone3	wrist x-y	-0.0948	52.3	[-0.3577, 0.1820]	0.550
csv_data_cannone3	shoulder-wrist y	0.4419	71.3	[0.2330, 0.6118]	<0.001
csv_data_cannone3	knee-ankle x	0.9568	39.9	[0.9193, 0.9771]	<0.001
csv_data_sikloski3	wrist x-y	-0.2892	64.2	[-0.4992, -0.0471]	0.026
csv_data_sikloski3	shoulder-wrist y	0.7808	65.2	[0.6634, 0.8607]	<0.001
csv_data_sikloski3	knee-ankle x	0.9632	44.3	[0.9333, 0.9798]	<0.001

The shared lower-body trajectory in Figure 1 may reflect both limb coordination and whole-body translation across the strip.

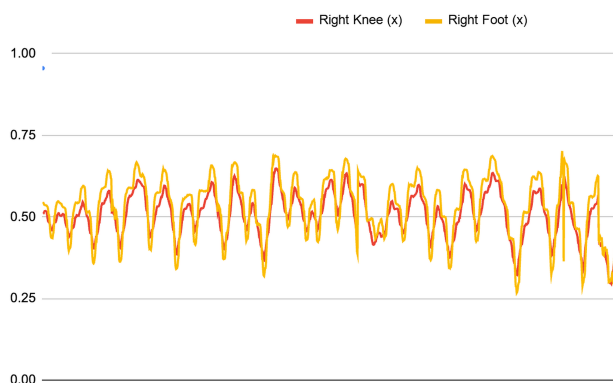


Figure. 1 Right knee and right foot index x-coordinates across the full csv_data_cannone4 clip (frames 0-1178), normalized to frame width.

Figure 2 shows the described correlations in graphical form and should be interpreted the same way with regard to the common motion problem. The effective sample sizes for these correlations ranged from 22.6 to 158.1, much smaller than the raw sample sizes of 1,153 to 1,396 frames. After correction, all four correlations between shoulders and wrists vertically, and all four correlations between knees and ankles horizontally resulted in confidence intervals that did not include zero and p-values less than 0.001 after Benjamini-Hochberg correction. For wrists x-y, only csv_data_sikloski3 showed an interval not including zero ($r = -0.2892$, 95% CI = $[-0.4992, -0.0471]$, $p = 0.026$), while the other three had adjusted p-values of 0.162, 0.667, and 0.550. Thus, the magnitude values represent effect size not statistical significance. Directional orientation of the wrist coordinates is given in Figure 3 below.

For the test video segment, the minimum MAE of 0.005991 was achieved using constant-velocity tracking. This is roughly 0.6% of the total image width and 1.5% of the total test motion distance. The persistence method returned a nearly similar MAE of 0.006163 and a better RMSE of 0.016547 compared to 0.024688 for constant velocity. Constant velocity performed slightly better on average whereas persistence had lesser missed detections. It was evident that both approaches outperformed the training-mean baseline which had a test MAE of 0.064874 and RMSE of 0.081646. These findings can be seen in Table 5.

The short term is illustrated in Figure 4 in two different perspectives. The top part shows that the single frame estimate closely follows the real trajectory of the wrist position. The bottom figure shows minor errors which are not noticeable due to overlapping lines, with the bigger bars representing those time periods when there was change in the wrist speed/direction.

Separate agreement was computed for both the workbook

Correlation Heatmap of Selected Right-Side Joints
Source: csv_data_cannone4, frames 0-1178

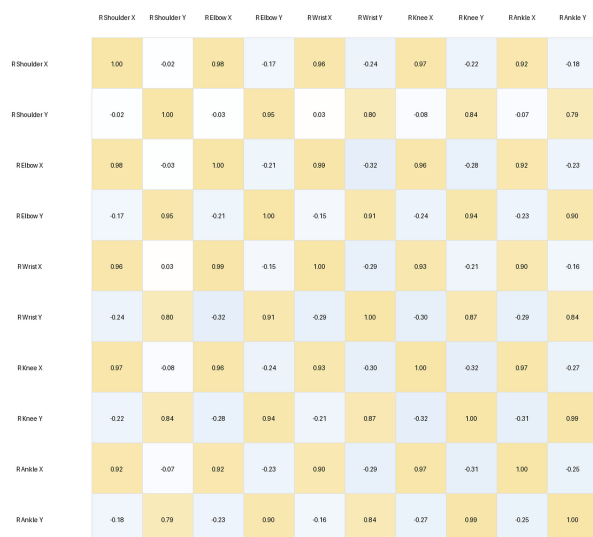


Figure. 2 Pearson-correlation heatmap of selected right-side joint coordinates for csv_data_cannone4, visualizing the correlations reported in Table 4.

clip and the complete video holdouts. The agreement for the 249 segments of the original clips with the four-class reviews was found to be 96.79% with a Cohen's Kappa of 0.9503. Agreement for the original validation clip was found to be 96.25% (Kappa = 0.9423), and the original test clip was 98.67% (Kappa = 0.9798). The 385 segments from the holdout videos in their entirety have a pre-adjudication agreement of 48.57% and Cohen's kappa of 0.3337, while validation and test agreements are 47.13% (kappa = 0.3200) and 49.56% (kappa = 0.3341), respectively. The reviewers have access to the shared candidate windows along with the proposed labels; hence, these numbers represent label agreement in the context of a rule-assisted review process. Disagreements have been settled via adjudication before evaluating the models. In the case where the model was trained only using the original segments, the performance was weak when evaluated on the complete video data set. After adding complete video validation labels to the training set, the macro F1 score for the current-segment MLP model on 227 test segments was 66.16%, versus 45.77% for the candidate-rule baseline model. The results are summarized in Table 6.

Figure 5 shows the confusion matrix for the held-out tests using the complete video test set for the current segment MLP network. The highest performing segments are the en garde and lunge/attack categories, while the retreat and transition/other segments tend to confuse each other more often. The cell in the matrix indicates the count and percentage of the segments in the action category. Figure 5 shows that

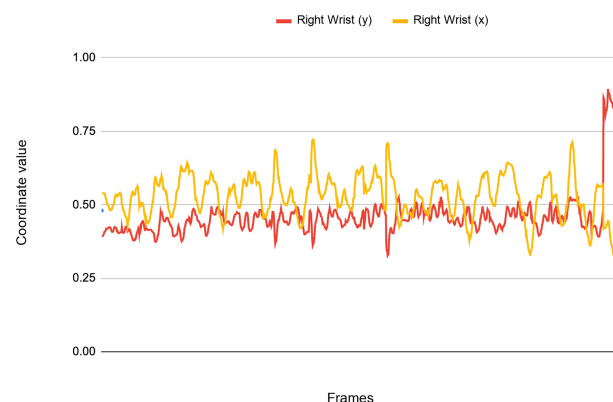


Figure. 3 Right wrist x- and y-coordinates across the full csv_data_cannone4 clip (frames 0-1178), the source trajectory for the coordinate-forecasting target.

Table 5 One-frame right-wrist x-coordinate forecasting errors with wrist $x(t+h)$, $h=1$ and Coordinate regression only

Model	Split	Pairs (N)	Motion range	MAE	RMSE
Training-mean	Train	2330	0.5122	0.052695	0.067570
Training-mean	Validation	1391	0.4089	0.055531	0.073770
Training-mean	Test	1395	0.3978	0.064874	0.081646
Persistence	Train	2330	0.5122	0.006585	0.014373
Persistence	Validation	1391	0.4089	0.007579	0.011366
Persistence	Test	1395	0.3978	0.006163	0.016547
Constant velocity	Train	2328	0.5122	0.006321	0.018069
Constant velocity	Validation	1390	0.4089	0.006925	0.011245
Constant velocity	Test	1394	0.3978	0.005991	0.024688

current-action performance varied by class rather than being uniformly distributed across the four-label taxonomy.

Figure 6 shows where correct classifications and errors occur across both held-out bouts while keeping each action segment as the evaluation unit. Table 7 reports aggregate and per-bout results for the validation-selected 30-sample upcoming-segment LSTM and compares its aggregate result with the current-segment MLP and candidate-rule baseline. The models address different classification tasks.

The validation-selected 30-sample LSTM correctly classified 64 of 225 complete-video test segments. It reached 37.89% accuracy, 35.30% balanced accuracy, and 27.45% macro F1 on MMN1tvPzNzM, and 21.54% accuracy, 21.85% balanced accuracy, and 18.60% macro F1 on rD9rm_MmAE4.

The chosen LSTM classifier predominantly classified transition/other and lunge/attack actions. The recall values for each action were 0% for en garde, 41.1% for lunge/attack, 17.8% for retreat, and 55.3% for transition/other. This reveals that the previous pose window approach involving only one athlete was not sufficient for classification of all the four ac-

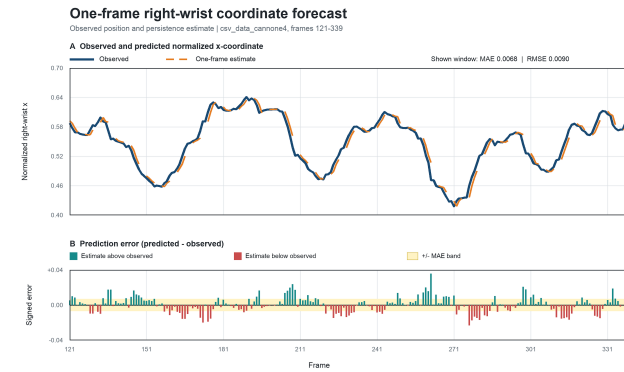


Figure. 4 One-frame right-wrist x-coordinate forecast for csv_data_cannone4, frames 121-339. Panel A: observed vs. persistence estimate. Panel B: signed error. This window (MAE 0.0068, RMSE 0.0090) is illustrative;

Table 6 Current-segment classification performance on complete-video holdouts.

Model	Split	Video ID	Segments	Accuracy	Balanced accuracy	Macro F1	Weighted F1
Current-segment MLP validation protocol	Original train	all	250	97.20%	98.62%	96.32%	97.26%
Current-segment MLP validation protocol	Complete-video validation	all	157	36.94%	41.63%	41.19%	37.02%
Current-segment MLP validation protocol	Complete-video test	all	227	34.36%	36.36%	35.04%	34.71%
Current-segment MLP final train plus validation	Complete-video test	all	227	68.28%	66.78%	66.16%	68.58%
Current-segment MLP final train plus validation	Complete-video test	MMN1tvPzNzM	96	75.00%	67.29%	67.09%	74.93%
Current-segment MLP final train plus validation	Complete-video test	rD9rm_MmAE4	131	63.36%	62.41%	61.68%	63.38%
Candidate-rule labeler	Complete-video test	all	227	45.81%	48.86%	45.77%	43.86%

tions. Predictions from the selected 30-sample LSTM were collected for all valid segments from the test videos. Each entry consists of the video name, segment interval, the true upcoming action label, the predicted label, probabilities of all the four labels, and whether the prediction was accurate or not.

Discussion

The last MLP classifier produced an accuracy of 68.28%, 66.78% balanced accuracy, and 66.16% macro F1 for 155 out of 227 video test segments. The rule candidate gave results with 45.81% accuracy and 45.77% macro F1. The greatest performance was demonstrated for the en garde and lunge/attack classes, while the poorest one was found for transition/other class; there were differences between test bouts. One frame ahead wrist prediction reached MAE of 0.005991 units, which means that simple rules allow predicting horizontal wrist movements one frame ahead. The main importance of the latter is determining the continuity measure. As far as the chosen model with 30 samples in an LSTM network is

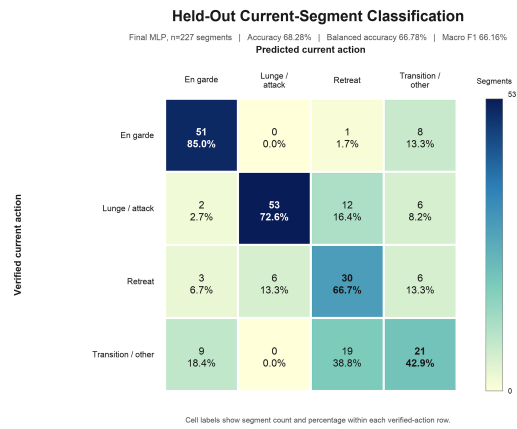


Figure. 5 Confusion matrix of current-segments, final MLP, 227 held-out test segments (MMN1tvPzNzM and rD9rm_MmAE4). Rows indicate true labels, while the columns denote predicted labels.

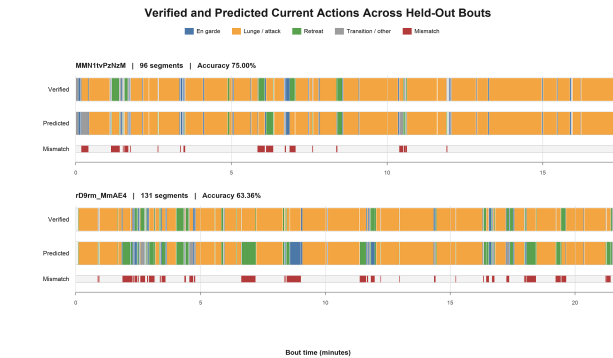


Figure. 6 Verified vs. predicted current-action timelines for the two held-out test bouts (~5 fps). Block width shows segment duration; the mismatch strip marks incorrect labels. Per-bout accuracy: 75.00% (MMN1tvPzNzM), 63.36% (rD9rm_MmAE4).

concerned, it showed the following results in the more difficult case of the upcoming segment: the accuracy was 28.44%, balanced accuracy was 28.55%, and the macro F1 value was 23.36%. From analyzing the confusion matrix and the per bout analysis, one can conclude that a single athlete pose trajectory did not distinguish between all the four possible upcoming actions. Generally, the results confirm the usefulness of pose trajectories and coordinate prediction, as well as the current action classification.

Limitations

In the current pilot work, there will be four processed épée clips for initial development as well as four full bouts for validation and testing purposes. Since there are just four video samples in total, each split at source level should be rather

Table 7 Comparison of upcoming- and current-segment classification.

Experiment	Task	Segments	Acc. (%)	Bal. acc. (%)	Macro F1 (%)
LSTM, all test bouts	Upcoming segment; selected 30-sample window	225	28.44	28.55	23.36
LSTM, MMN1tvPzNzM	Upcoming segment; test bout	95	37.89	35.30	27.45
LSTM, rD9rm_MmAE4	Upcoming segment; test bout	130	21.54	21.85	18.60
Current-segment MLP	Current segment; pose summaries	227	68.28	66.78	66.16
Candidate-rule baseline	Candidate vs. adjudicated current label	227	45.81	48.86	45.77

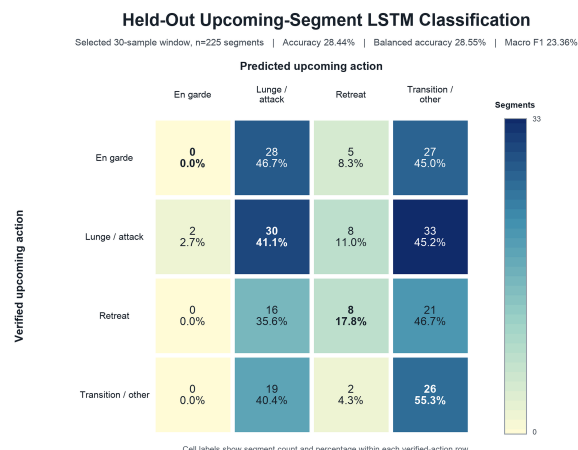


Figure. 7 Upcoming-segment confusion matrix selected 30-sample-window LSTM, 225 held-out test segments (MMN1tvPzNzM and rD9rm_MmAE4). Rows are verified labels, columns are predictions; cells show count and row percentage.

coarse. A split at ratio 2/2 results in having two test bouts, while the split at 3/1 ratio gives extra data for development but has only one test bout. A segment-based random split has not been used because it could lead to the appearance of similar images from the same video in both training and testing datasets. The original workbook does not have the exact correspondence between each worksheet and video window. However, the analysis process for the single pose case is confined to just one stream of observed body poses. It does not consider the distance to the opponent, the position of the sword, the instigator of the action, or any reaction whatsoever. In addition to that, variations can occur because of different cameras, recaptures, officials, occlusion, and tracking. The conflicts between the annotators in the overlapping candidate windows have been resolved with the help of rules-based arbitration.

Future Work

Further research should include named athletes and two fencers, features of the blade, non-match filtering and replay filtering, and independently defined time bounds. There

should be a groupwise leave-one-video-out sensitivity analysis where three bouts are trained and the fourth tested in each fold, using out-of-fold and per-bout evaluations; however, a larger number of data is required to maintain a completely unused test set. The increased size of the held-out data can facilitate comparison of the performance of recurrent neural networks, convolutional neural networks, transformers, and interaction-aware networks on the same data splits. Interaction-aware networks can model the response to a thrust of the opponent, as done by Honda et al.¹⁶.

Conclusion

Three bounded findings can be derived from the pilot study. First, normalized trajectories of pose coordinates can serve as an adequate description of motion patterns of visible joints in the fencing videos. Second, single frame-based forecasting of wrist position is mostly associated with local consistency of the coordinate motion because the lowest test MAE was 0.005991 normalized units (or approximately 0.6% of the image width). Third, the current-segment multilayer perceptron achieved 68.28% classification accuracy and 66.16% macro F1 score on the held-out actions, whereas the selected 30 samples of the validation-based long short-term memory obtained 28.44% classification accuracy and 23.36% macro F1 score on the more difficult task of predicting the next action.

References

- 1 V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann. BlazePose: On-device real-time body pose tracking. arXiv:2006.10204. 2020, <https://doi.org/10.48550/arXiv.2006.10204>.
- 2 C. S. T. Hii, K. B. Gan, N. Zainal, N. M. Ibrahim, S. Azmin, S. H. M. Desa, B. van de Warrenburg, H. W. You. Automated gait analysis based on a marker-free pose estimation model. Sensors. Vol. 23(14), pg. 6489, 2023, <https://doi.org/10.3390/s23146489>.
- 3 K. Chida, T. Inami, S. Yamaguchi, T. Nishioka, Y. Yoshida, N. Kohtake. Assessing the validity of two-dimensional video analysis for measuring lower limb joint angles during fencing lunge. Frontiers in Sports and Active Living. Vol. 6, pg. 1335272, 2024, <https://doi.org/10.3389/fspor.2024.1335272>.
- 4 D. Pagnon, M. Domalain, L. Reveret. Pose2Sim: An end-to-end workflow for 3D markerless sports kinematics-Part 1: Robustness. Sensors. Vol. 21(19), pg. 6530, 2021, <https://doi.org/10.3390/s21196530>.
- 5 N. Nakano, T. Sakura, K. Ueda, L. Omura, A. Kimura, Y. Iino, S. Fukushima, S. Yoshioka. Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. Frontiers in Sports and Active Living. Vol. 2, pg. 50, 2020, <https://doi.org/10.3389/fspor.2020.00050>.
- 6 N. J. Cronin, J. Walker, C. B. Tucker, G. Nicholson, M. Cooke, S. Merlino, A. Bissas. Feasibility of OpenPose markerless motion analysis in a real athletics competition. Frontiers in Sports and Active Living. Vol. 5, pg. 1298003, 2024, <https://doi.org/10.3389/fspor.2023.1298003>.

- 7 L. Needham, M. Evans, L. Wade, D. P. Cosker, M. P. McGuigan, J. L. Bilzon, S. L. Colyer. The development and evaluation of a fully automated markerless motion capture workflow. *Journal of Biomechanics*. Vol. 144, pg. 111338, 2022, <https://doi.org/10.1016/j.jbiomech.2022.111338>.
- 8 C. Mercadal-Baudart, C.-J. Liu, G. Farrell, M. Boyne, J. Gonzalez Escribano, A. Smolic, C. Simms. Exercise quantification from single camera view markerless 3D pose estimation. *Heliyon*. Vol. 10(6), pg. e27596, 2024, <https://doi.org/10.1016/j.heliyon.2024.e27596>.
- 9 M. Mundt, S. Colyer, L. Wade, L. Needham, M. Evans, E. Millett, J. Alderson. Automating video-based two-dimensional motion analysis in sport? Implications for gait event detection, pose estimation, and performance parameter analysis. *Scandinavian Journal of Medicine & Science in Sports*. Vol. 34(7), pg. e14693, 2024, <https://doi.org/10.1111/sms.14693>.
- 10 C. Magnani, E. Defrasne Ait-Said. Geometrical analysis of motion schemes on fencing experts from competition videos. *PLOS ONE*. Vol. 16(12), pg. e0261888, 2021, <https://doi.org/10.1371/journal.pone.0261888>.
- 11 Y. Guan, L. Guo, N. Wu, L. Zhang, D. E. R. Warburton. Biomechanical insights into the determinants of speed in the fencing lunge. *European Journal of Sport Science*. Vol. 18(2), pg. 201–208, 2018, <https://doi.org/10.1080/17461391.2017.1414886>.
- 12 L. Bottoms, A. Greenhalgh, J. Sinclair. Kinematic determinants of weapon velocity during the fencing lunge in experienced epee fencers. *Acta of Bioengineering and Biomechanics*. Vol. 15(4), pg. 109–113, 2013, <https://pubmed.ncbi.nlm.nih.gov/24479483/>.
- 13 A. Akbas, W. Marszalek, B. Bacik, G. Juras. Two aspects of feedforward control during a fencing lunge: Early and anticipatory postural adjustments. *Frontiers in Human Neuroscience*. Vol. 15, pg. 638675, 2021, <https://doi.org/10.3389/fnhum.2021.638675>.
- 14 A. Akbas, W. Marszalek, A. Brachman, G. Juras. Influence of target width and distance on postural adjustments in a fencing lunge. *Journal of Human Kinetics*. Vol. 87, pg. 35–45, 2023, <https://doi.org/10.5114/jhk/161572>.
- 15 Z. Borysiuk, M. Konieczny, M. Blaszczyzyn, W. Blach, Z. Obminski. The phenomenon of anticipation in fencing: An applicability approach. *Frontiers in Sports and Active Living*. Vol. 6, pg. 1387013, 2024, <https://doi.org/10.3389/fspor.2024.1387013>.
- 16 Y. Honda, R. Kawakami, T. Naemura. RNN-based motion prediction in competitive fencing considering interaction between players. In *Proceedings of the British Machine Vision Conference 2020*. <https://open-vision.sc.e.titech.ac.jp/~reikawa/publication/honda-bmvc2020.pdf>.
- 17 E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pg. 6992–7001, 2020, https://openaccess.thecvf.com/content_CVPR2020/html/Corona_Context-Aware_Human_Motion_Prediction_CVPR2020_paper.html.
- 18 F. Malawski, B. Kwolek. Recognition of action dynamics in fencing using multimodal cues. *Image and Vision Computing*. Vol. 75, pg. 1–10, 2018, <https://doi.org/10.1016/j.imavis.2018.04.005>.
- 19 F. Malawski, M. Krupa. Temporal segmentation of actions in fencing footwork training. *Computer Science Research Notes*. Vol. 33, pg. 242–248, 2023, <https://doi.org/10.24132/CSRN.3301.28>.
- 20 K. Zhu, A. Wong, J. McPhee. FenceNet: Fine-grained footwork recognition in fencing. *arXiv:2204.09434*. 2022, <https://doi.org/10.48550/arXiv.2204.09434>.
- 21 C. Lai, J. Mo, H. Xia, Y.-F. Wang. FACTS: Fine-grained action classification for tactical sports. *arXiv:2412.16454*. 2024, <https://doi.org/10.48550/arXiv.2412.16454>.
- 22 Y. He, Z. Yuan, Y. Wu, L. Cheng, D. Deng, Y. Wu. ViSTec: Video modeling for sports technique recognition and tactical analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38(8), pg. 8490–8498, 2024, <https://doi.org/10.1609/aaai.v38i8.28692>.
- 23 H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pg. 2969–2978, 2022, https://openaccess.thecvf.com/content_CVPR2022/html/Duan_Revisiting_Skeleton-Based_Action_Recognition_CVPR2022_paper.html.
- 24 D. Shao, Y. Zhao, B. Dai, D. Lin. FineGym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pg. 2616–2625, 2020, https://openaccess.thecvf.com/content_CVPR2020/html/Shao_FineGym_A_Hierarchical_Video_Dataset_for_Fine-Grained_Action_Understanding_CVPR2020_paper.html.
- 25 M. Yan, Y. Zhang, S. Cai, S. Fan, X. Lin, Y. Dai, S. Shen, C. Wen, L. Xu, Y. Ma, C. Wang. RELI11D: A comprehensive multimodal human motion dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pg. 2250–2262, 2024, https://openaccess.thecvf.com/content_CVPR2024/html/Yan_RELI11D_A_Comprehensive_Multimodal_Human_Motion_Dataset_and_Method_CVPR2024_paper.html.
- 26 K. Chida, T. Inami, S. Yamaguchi, Y. Yoshida, N. Kohtake. Interaction of target distance and movement type on lower limb kinematics in fencing. *Frontiers in Sports and Active Living*. Vol. 8, pg. 1756804, 2026, <https://doi.org/10.3389/fspor.2026.1756804>.
- 27 YouTube. Source fencing bout video used for pose extraction. <https://www.youtube.com/watch?v=MVqBp6dDTXg>, n.d.
- 28 YouTube. Source fencing bout video used for pose extraction. <https://www.youtube.com/watch?v=1CP-lrvsR44>, n.d.
- 29 YouTube. Source fencing bout video used for pose extraction. <https://www.youtube.com/watch?v=Dn8-M2zW9PU>, n.d.
- 30 YouTube. Budapest World Cup 2026 MS - L64 - Aron Szilagyi HUN v Cosimo Bertini ITA. <https://www.youtube.com/watch?v=aVtn119I1U4>, n.d.
- 31 YouTube. Budapest World Cup 2026 SMS - L16 - Oh Sanguk KOR v Remi Garrigue FRA. <https://www.youtube.com/watch?v=UO90aqHcAlU>, n.d.
- 32 YouTube. Turin Grand Prix 2026 SMF - GOLD - Nick Itkin USA vs Guillaume Bianchi ITA. <https://www.youtube.com/watch?v=MMNltvPzNzM>, n.d.
- 33 YouTube. CIP World Cup 2026 SMF - GOLD - Ryan Choi HKG v Alex Massialas USA. https://www.youtube.com/watch?v=rD9rm_MmAE4, n.d.
- 34 Z. Liu, P. Su, S. Wu, X. Shen, H. Chen, Y. Hao, M. Wang. Motion prediction using trajectory cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pg. 13299–13308, 2021, https://openaccess.thecvf.com/content_ICCV2021/html/Liu_Motion_Prediction_Using_Trajectory_Cues_ICCV2021_paper.html.