SyncScore: A Framework for Synchronization Scoring in Group Sports via Human Pose Estimation

Khai Pin Ang*, Iven Zi Yin Low*, Yumun Hooi*, Yuen Peng Loh*†
† Centre for Image and Vision Computing (CIVC), CoE for Artificial Intelligence
* Faculty of Computing and Informatics, Multimedia University
Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
E-mail: {1211101248@student.,yploh@}mmu.edu.my

***Abstract*—Human Pose Estimation (HPE) plays a critical role in performance analysis for synchronized sports such as Taekwondo Poomsae, where timing and coordination between participants are essential. Traditional synchronization scoring methods rely on manual expert judgment, which can be subjective and inconsistent. This paper proposes an automated framework for synchronization scoring, leveraging the ViTPose model for high-precision keypoint detection and robust generalization. To estimate synchronization scores, we compute inter-person landmark distances using Euclidean distance applied to aligned and temporally smoothed 2D landmarks. These distances are then used as input features to regression models that estimate synchronization scores aligned with expert assessments. The framework was validated using the benchmark LSP dataset for HPE evaluation and a custom-annotated Taekwondo Poomsae dataset comprising 70 videos. Experimental results show that ViTPose, fine-tuned on the custom dataset provided good performance for complex HPE. Among the regression approaches, the combination of ViTPose-L with Support Vector Regression (SVR) achieved the highest synchronization score estimation with an R^2 value of 0.3915. These findings demonstrate the potential of the proposed framework for objective and precise scoring in synchronized sports performance.**

I. Introduction

Human Pose Estimation (HPE), a vital branch of computer vision, involves detecting anatomical landmarks or keypoints on the human body from images or videos. Keypoints such as elbows, wrists, shoulders, and knees allow accurate tracking of body movements in real time [1]. HPE has found diverse applications across domains including sports analytics, physical therapy, and e-commerce [2]. In synchronized group sports like Karate Kata and Taekwondo Poomsae, precise coordination and alignment of multiple athletes' movements are critical for performance and competition. However, assessing synchronization traditionally relies on manual expert evaluation, which can be subjective, inconsistent, and prone to biases [3] making accurate, objective synchronization scoring a challenging task.

The advancement of HPE models offers significant potential for automating sports assessments, enabling detailed analysis of posture, movement, and coordination. While HPE has been extensively applied to single-athlete pose estimation and general sports analytics [4], [5], its application to synchronized group sports, particularly for scoring and evaluating synchronicity remains underexplored. Automating synchronization scoring can reduce human bias and provide precise, real-time feedback to athletes, coaches, and judging panels, supporting better training outcomes and performance evaluations. Despite the growing sophistication of HPE models, challenges persist in accurately detecting complex, dynamic, and occluded poses, in particular for a domain such as martial arts and other synchronized sports.

This paper proposes a novel framework for synchronization analysis in group sports using HPE, the SyncScore .In particular we focus on synchronized two-person Teakwondo Poomsae where two athletes perform these forms in unison, and scoring is based on both individual technical execution and synchronization between performers. We specifically address the synchronization scoring aspect of the Poomsae performance by exploring the ViTPose [6] to detect and track athlete landmarks. Our approach involves extracting, smoothing, and aligning these landmarks to compute similarity measures, i.e. Euclidean distance, that quantify synchronization. These distances are then mapped to synchronization scores using regression models, with Support Vector Regressor delivering the best performance. We validate our framework on a self-collected, annotated Taekwondo Poomsae dataset, demonstrating that ViTPose outperforms a baseline MoveNet [7] in both pose accuracy and synchronization scoring estimation. Our work offers an objective, precise alternative to manual scoring, providing valuable tools for athlete evaluation, coaching, and competition preparation in synchronized group sports.

II. Related Work

A. General Human Pose Estimation

General HPE involves detecting anatomical keypoints such as shoulders, elbows, and knees from images or videos. With the advent of deep learning, HPE has significantly advanced, especially in addressing challenges like occlusion, scale variation, and dynamic motion. Early approaches such as DeepPose [8] framed HPE as a regression problem using deep convolutional networks. OpenPose [9] and PifPaf [10] later introduced bottom-up methods capable of multi-person detection using part affinity fields and composite fields, respectively. OmniPose [11] improved accuracy further by integrating HRNet [1] with Gaussian heatmap modulation.

Subsequently, the MoveNet, developed by Google, is a fast and accurate bottom-up model optimized for real-time

979-8-3315-7206-8/25/$31.00

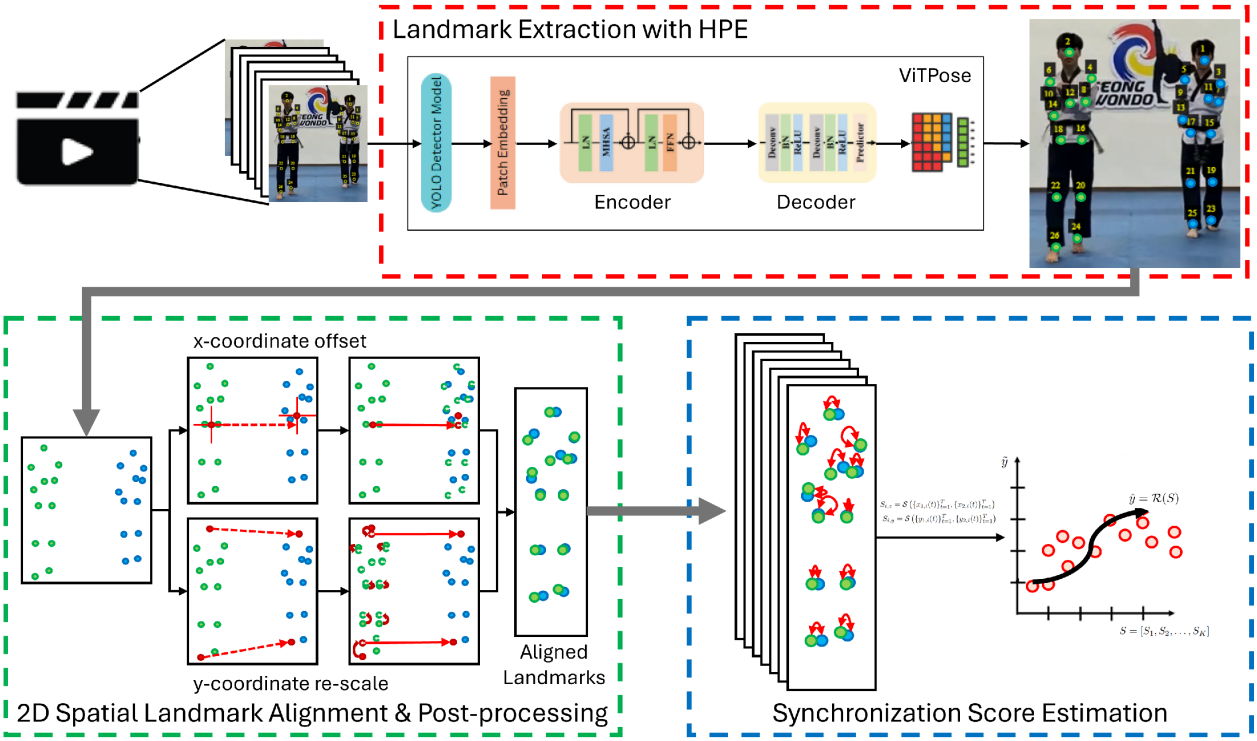

Fig. 1. Overview of proposed framework, SyncScore.

applications [7]. It uses a MobileNetV2 [12] backbone with a feature pyramid network and four prediction heads to detect person centers and keypoints. On the other hand, BlazePose also targets real-time pose tracking on mobile devices [13], employing an encoder-decoder design that combines heatmap generation with coordinate regression. The heatmap branch is discarded during inference for improved efficiency.

Following that, transformer-based models have recently reshaped the HPE landscape. ViTPose [6] stands out by using a plain, non-hierarchical Vision Transformer backbone without reliance on CNNs. Pretrained via masked image modeling, ViTPose pairs this backbone with a lightweight decoder to generate pose heatmaps. Even though the architecture is structurally simple, ViTPose achieves state-of-the-art results. While these general-purpose HPE models have shown strong performance on standard benchmarks, their potential remains underutilized in synchronized sports assessment, where precise evaluation of coordinated, high-speed group movements is essential. Hence, this paper explores the application of HPE for scoring synchronization, addressing the gap by proposing a framework tailored for performance analysis in group sports.

B. Human Pose Estimation in Sports

Since the start, HPE has been of interest to sports for applications such as performance analysis, movement correction, and automated coaching. Foundational work like DeepPose [8] enabled CNN-based keypoint regression, leading to systems such as the AI Coach [4], which combined visual tracking and pose modeling for personalized feedback. This system demonstrated effectiveness on a Freestyle Skiing dataset by incorporating temporal pose correction and visual suggestions. Similarly, HPE has been applied in martial arts; [14] proposed a CNN-based system for static pose evaluation using domain-specific metrics like joint angles and keypoint displacement, though it struggled with occlusions and motion blur. Complementing vision-based approaches, [3] introduced a multi-modal method using wearable IMUs and wavelet transforms for joint angle tracking, achieving high agreement with expert scores albeit with the drawback of requiring specialized hardware.

More recently, [5] critically evaluated BlazePose and OpenPose for sports-specific tasks, showing that general-purpose HPE models often fail to capture fine-grained details like wrist and foot locations during dynamic exercises. Their findings highlight the importance of domain adaptation when applying HPE in real-world sports settings. Building on these insights, our work addresses a key gap in synchronized group sports, an area less explored in prior research. We propose a framework that captures inter-person pose similarity and synchronization in martial arts routines, enabling automated scoring and team performance analysis in multi-person sports contexts.

III. Proposed Framework

As a proof-of-concept, our proposed SyncScore framework focuses on two-person Taekwondo Poomsae performance. The framework starts with (1) human pose estimation to extract body landmarks from video frames, followed by (2) 2D spatial landmark coordinates alignment of both participants.Then, (3) erroneous or unstable keypoints are handle through post-

processing and subsequently (4) inter-person landmark distances are calculated across corresponding frames to capture the level of synchronization. These extracted features are finally used to (5) train a machine learning model that predicts an overall performance score, enabling automated assessment of synchronized movement quality. Figure 1 shows the overview of the proposed SyncScore.

A. Landmark Extraction with HPE

Motivated by the developments of HPE algorithsm, this study explores the promising ViTPose [6] for their applicability in detecting martial arts pose landmarks and validating the proposed synchronized sports scoring framework.

ViTPose adopts a top-down pose estimation approach, first using a YOLO object detector to localize individuals within each video frame. The detected person regions are then cropped and passed through a Vision Transformer (ViT) [15] backbone, which encodes the input into tokenized representations. These features are decoded into heatmaps representing keypoint locations using a lightweight decoder.

Specifically in this work, the HPE model is adapted to extract 13 keypoints from each participants instead of the common 17 each (eyes and ears are removed as they do not contribute to the poomsae poses). Therefore, each given frame of a Poomsae performance will extract 26 keypoints, assigned to the respective participant. Since each participant may appear at different distances, their keypoints can vary in scale, causing one performer to appear taller or larger than the other, while also being located at different positions within the frame. These spatial and scale inconsistencies make direct comparison between the two sets of keypoints unreliable. To address this, the next stage performs 2D coordinate normalization, aligning both participants to a consistent reference frame to enable the keypoint positional differences to be computed.

B. 2D Spatial Alignment of Landmark Coordinates

To enable fair and accurate comparison of movements between participants in synchronized Poomsae performances, normalization of landmark coordinates in the 2D frame space is essential. Due to differences in camera perspective, participants may appear at different positions and scales within the frame. These discrepancies can lead to misleading interpretations of synchrony if raw coordinates are used directly.

1) X-Coordinate Offset Adjustment: Participants often begin at different horizontal positions in the frame, resulting in significant variation in X-axis landmark coordinates. This spatial discrepancy must be corrected to ensure alignment. Consider the set of participant landmarks is represented as $x_m = \{x_{m,1}, x_{m,2}, ..., x_{m,13}\}$ where m is the participant ID. An offset, Δx is applied to the X coordinates of participant 1 by calculating the difference in mean X values, μ_x between the two participants, $\Delta x = \mu_{x2} - \mu_{x1}$ where μ_{x1} and μ_{x2} are the mean X coordinates of participant 1 and 2, respectively. This offset Δx is then added to all of participant 1's X coordinates, $x'_{1,i} = x_{1,i} + \Delta x$. This operation aligns participant 1's horizontal pose positions with participant 2's, enabling proper spatial comparisons.

2) Y-Coordinate Re-scaling: Vertical landmark coordinates (Y-axis) are affected by participants' varying heights and distances from the camera, which can distort the scale and position of landmarks. Applying a simple offset is insufficient for accurate vertical alignment in this case. Instead, we rescale the Y coordinates of one participant to match the other by using reference landmarks that represent the participant's effective height. Specifically, the nose ($y_{m,\text{nose}}$) and right ankle landmark ($y_{m,\text{ankle}}$). Hence, for each participant $m \in \{1, 2\}$, we define the vertical height range as

$$H_m = y_{m,\text{ankle}} - y_{m,\text{nose}} \tag{1}$$

We then normalize the Y coordinates of participant 1 relative to participant 2 by applying the following transformation:

$$y'_{1,i} = \frac{y_{1,i} - y_{1,\text{nose}}}{H_1} \times H_2 + y_{2,\text{nose}} \tag{2}$$

where $y_{1,i}$ is the original Y coordinate of the i-th landmark for participant 1, and $y'_{1,i}$ is the rescaled Y coordinate aligned to participant 2's vertical range. This operation scales and shifts participant 1's vertical landmarks to the same height range and position as participant 2, enabling more accurate vertical comparisons between the two participants that is invariant to height and perspective differences.

Together, these coordinate re-scaling techniques effectively align the landmark data of the two participants in the 2D frame, accounting for differences in horizontal positioning, height, and camera perspective. This alignment ensures that the spatial features extracted from each performer are directly comparable for synchronization computation.

C. Landmark Post-processing

During landmark detection, pose estimation models may produce erroneous landmarks, often appearing as sharp spikes or sudden deviations in coordinate trajectories. These outliers can distort movement analysis and compromise synchronization computation. To mitigate this, median smoothing is applied using a sliding window of size 5, replacing each coordinate with the median of its temporal neighbors (neighboring video frames). This non-linear filtering effectively suppresses outliers while preserving landmark positions, as illustrated with the sample signals shown in Fig. 2. On the other hand, there may be cases where the landmarks are not detected by the models at a specific frame. Hence, we fill in the missing landmarks with those detected in the previous frame. As a whole, these processes enhances the accuracy, reliability, and overall robustness of the synchronization evaluation.

After normalization and temporal smoothing of landmark positions, the next step is to quantify synchronization between participants by comparing the movement of corresponding landmarks over time. This is done by computing distances on the temporal sequences of 2D coordinates.

Let $L_{m,i}(t) = (x_{m,i}(t), y_{m,i}(t))$ represent the 2D coordinates of the i-th landmark for participant m (where

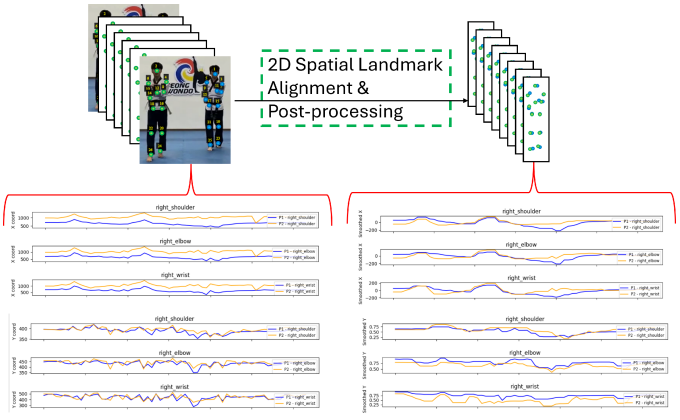

Fig. 2. X- and Y-coordinates of landmarks before (left) and after (right) 2D alignment and post processing.

$m \in \{1, 2\}$) at frame t. Here, $i = 1, \ldots, K$ landmarks, and $t = 1, \ldots, T$ frames. For each landmark i, similarity scores are computed separately for the X and Y coordinates using a general similarity function $\mathcal{S}$, where

$$S_{i,x} = \mathcal{S}\left(\{x_{1,i}(t)\}_{t=1}^{T}, \{x_{2,i}(t)\}_{t=1}^{T}\right); \quad (3)$$

$$S_{i,y} = \mathcal{S}\left(\{y_{1,i}(t)\}_{t=1}^{T}, \{y_{2,i}(t)\}_{t=1}^{T}\right) \quad (4)$$

The similarity function $\mathcal{S}$ can be any suitable measure such as histogram intersection, Euclidean distance, etc.. The combined similarity scores for all landmarks in both X and Y dimensions,$\{S_{i,x}, S_{i,y}\}_{i=1}^{K}$, form a feature vector representing the inter-person motion similarity profile across both horizontal and vertical landmark movements.

D. Synchronization Score Estimation

Following the formation of the similarity scores for each landmark, they are then used collectively as features to predict the overall synchronization score. For each landmark i, the similarity score S_i is a sum of the similarity measures computed separately over the X and Y coordinate sequences,

$$S_i = S_{i,x} + S_{i,y} \quad (5)$$

where $S_{i,x}$ and $S_{i,y}$ denote the similarity values for the X and Y trajectories of the i-th landmark, respectively. This results in a feature vector $S = [S_1, S_2, \ldots, S_K] \in \mathbb{R}^K$ where $K = 13$ is the number of landmarks. The entire vector S therefore serves as the input to the regression model $\mathcal{R}(\cdot)$, which predicts the synchronization performance score, $\hat{y} = \mathcal{R}(S)$.

The regression model is trained on synchronization scores provided by human experts, $\tilde{y}$ used as ground truth labels. These scores follow criteria published by the World Taekwondo Federation (WTF) [16], where poomsae performance is evaluated based on accuracy and presentation, with presentation specifically reflecting rhythm and tempo, key indicators of synchronization between participants. Hence, the model effectively maps the computed landmark distances to human-perceived synchronization levels.

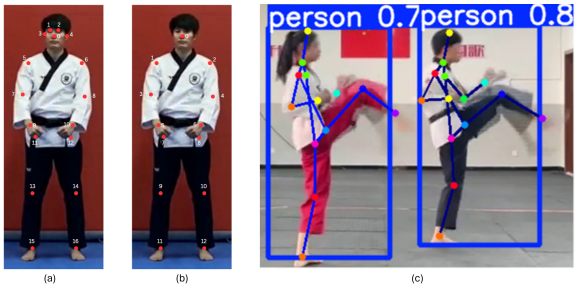

Fig. 3. (a) Landmarks obtained from original MoveNet model on self-collected 2-person Taekwondo Poomsae performance videos, and (b) Landmarks manually corrected for synchronization assessment, include of the include nose, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, and left ankle. (c) Sample landmark detection by ViTPose on complex poses.

IV. Experiments and Results

A. Datasets

For the experiments in this study, two datasets were utilized, the benchmark Leeds Sports Pose (LSP) dataset [17] to evaluate and select the most suitable HPE model based on landmark detection performance, and a self-collected 2-person Taekwondo Poomsae video dataset annotated with 13 keypoints per person used for synchronization analysis..

The ground truth annotations for the Taekwondo dataset keypoints were obtained by initially predicted using the pre-trained MoveNet model providing 17 keypoints, followed by manual correction using the VGG Image Annotator (VIA) tool [18], [19] to obtain the final keypoints covering the key body landmarks for Poomsae performance, as shown in Fig. 3 (a) and (b).

A total of 70 Taekwondo Poomsae videos were used, from which 60 videos were allocated for training and 10 for testing. From the training set, 400 representative frames were selected and manually annotated, while an additional 100 annotated frames were obtained from the testing set. This process resulted in a high-quality custom dataset of 500 annotated frames, supporting the evaluation of pose estimation accuracy and synchronization performance.

B. Implementation Details

For implementation, we fine-tuned three ViTPose model variants (base, large, and huge) on two datasets: a custom Taekwondo Poomsae dataset and the benchmark LSP dataset. To evaluate model performance under different training conditions, we experimented with seven keypoint regression loss functions to train the model: Mean Squared Error (MSE), Mean Absolute Error (MAE), Smooth Mean Absolute Error (SMAE), Combined Target MSE Loss (CTML) [20], Online Hard Keypoint Mining MSE (OHKM) [21], Huber, and Kullback–Leibler Divergence Loss (KLD). Each model was pre-trained using the COCO dataset [22], then fine-tuned for 30 epochs with a learning rate of 0.0005 with the loss functions varied to assess their impact on landmark estimation accuracy.

TABLE I
HPE EVALUATION OF VITPOSE ON LSP DATASET

Size	Loss	AP@0.5	AP@0.75	mAP	AR@0.5	AR@0.75	mAR
Base	MSE	0.703	**0.523**	0.522	0.829	0.707	0.703
	CTML	**0.716**	0.520	**0.531**	**0.835**	**0.707**	**0.706**
	Huber	0.701	0.519	0.528	0.832	0.707	0.705
	KLD	0.364	0.134	0.167	0.595	0.352	0.371
	MAE	0.654	0.507	0.507	0.800	0.698	0.687
	OHKM	0.703	0.523	0.522	0.829	0.707	0.697
	SMAE	0.701	0.519	0.528	0.827	0.703	0.701
Large	MSE	0.678	0.552	0.540	0.816	0.731	0.716
	CTML	**0.696**	0.542	0.527	**0.825**	0.728	0.707
	Huber	0.676	0.554	0.542	0.814	0.739	**0.740**
	KLD	0.416	0.163	0.207	0.639	0.397	0.419
	MAE	0.694	**0.581**	**0.565**	0.823	**0.749**	0.731
	OHKM	0.675	0.541	0.536	0.812	0.728	0.714
	SMAE	0.676	0.541	0.536	0.814	0.728	0.715
Huge	MSE	**0.673**	0.523	0.514	0.810	0.713	0.699
	CTML	0.629	0.472	0.473	0.787	0.679	0.671
	Huber	0.673	0.523	0.512	0.810	0.711	0.698
	KLD	0.387	0.152	0.190	0.618	0.386	0.407
	MAE	0.672	**0.524**	**0.520**	**0.812**	**0.717**	**0.706**
	OHKM	0.673	0.525	0.513	0.810	0.713	0.699
	SMAE	0.673	0.523	0.512	0.810	0.711	0.698

TABLE II
HPE EVALUATION OF VITPOSE ON TAEKWONDO POOMSAE DATASET

Size	Loss	AP@0.5	AP@0.75	mAP	AR@0.5	AR@0.75	mAR
Base	MSE	0.945	0.820	0.768	0.955	0.875	**0.841**
	CTML	**0.948**	**0.830**	**0.773**	**0.960**	**0.880**	0.831
	Huber	0.945	0.826	0.768	0.960	0.880	0.831
	KLD	0.916	0.809	0.696	0.955	0.865	0.799
	MAE	0.932	0.820	0.745	0.955	0.880	0.811
	OHKM	0.945	0.828	0.770	0.960	0.880	0.832
	SMAE	0.945	0.826	0.768	0.960	0.885	0.831
Large	MSE	0.930	0.822	0.771	0.955	0.885	**0.837**
	CTML	**0.933**	**0.826**	**0.774**	**0.955**	**0.880**	0.833
	Huber	0.930	0.821	0.770	0.955	0.885	0.837
	KLD	0.910	0.801	0.698	0.955	0.885	0.805
	MAE	0.925	0.810	0.753	0.955	0.885	0.824
	OHKM	0.930	0.822	0.772	0.955	0.880	0.837
	SMAE	0.930	0.821	0.770	0.955	0.885	0.837
Huge	MSE	0.930	0.837	0.784	0.955	0.890	0.847
	CTML	**0.932**	**0.837**	**0.787**	**0.955**	**0.890**	**0.848**
	Huber	0.930	0.837	0.784	0.955	0.890	0.848
	KLD	0.920	0.825	0.738	0.955	0.890	0.829
	MAE	0.931	0.828	0.774	0.955	0.890	0.842
	OHKM	0.930	0.837	0.785	0.955	0.890	0.848
	SMAE	0.930	0.837	0.784	0.955	0.890	0.847

C. HPE Evaluation

The quantitative results of 3 ViTPose model variants (Base, Large, Huge) experiments are summarized in Table I for the LSP benchmark dataset and Table II for the Taekwondo Poomsae dataset. Both tables present standard COCO evaluation metrics (AP@0.5, AP@0.75, mAP, AR@0.5, AR@0.75, mAR) across different loss functions and model sizes.

In Table I, the MAE loss function delivered the best performance for both the large and huge ViTPose variants, achieving a maximum mAP of 0.565. This demonstrates a clear improvement over the originally implemented MSE loss on the LSP dataset. Among all losses tested, KLD consistently underperformed, highlighting its limitations for keypoint heatmap regression. This shows the strong generalization capabilities of MAE and CTML losses for pose estimation in generic sports.

Table II presents results on the Taekwondo Poomsae dataset, where CTML emerged as the most effective loss function across all model sizes. The highest mAP of 0.787 was obtained using the ViTPose-huge model. OHKM and Huber losses also yielded competitive results, while KLD again showed the weakest performance. Notably, CTML outperformed MSE, the original baseline loss used in ViTPose, confirming its suitability for fine-grained martial arts pose estimation. Overall, the evaluation results demonstrate that MAE and CTML are most suited for sports pose estimation that often contains challenging poses, as shown in Fig. 3.

TABLE III
SYNCHRONIZATION SCORE ESTIMATION

Model	Loss	Regressor	R^2	MSE	MAE
MoveNet	CCE	ElasticNet	0.3545	0.0039	0.0509
ViTPose-B	MSE	Ridge	0.2830	0.0043	0.0522
	CTML	Ridge	0.2825	0.0043	0.0524
ViTPose-L	MSE	SVR	**0.3915**	**0.0036**	**0.0473**
	CTML	Ridge	0.2839	0.0043	0.0524
ViTPose-H	MSE	SVR	0.3842	0.0037	0.0488
	CTML	Ridge	0.2843	0.0043	0.0524

D. Synchronization Scoring Evaluation

To assess the effectiveness of ViTPose in synchronized sports scoring, we used the Euclidean distance between corresponding keypoints as the input distance measure to evaluate the landmark differences of the two participants in the poomsae videos. These distances which are aggregated over all 13 keypoints, serves as feature vectors. In our experiments, we employed the Support Vector Regressor (SVR) and Ridge regressor as the models for synchronization scores estimation. Table III presents the performance metrics, including R-Squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE), for ViTPose variants trained with different loss functions, alongside a comparison with a COCO pre-trained MoveNet with categorial cross entropy (CCE) and ElasticNet.

Among all configurations, the ViTPose-Large model trained with MSE loss and evaluated using Support Vector Regression (SVR) achieved the highest R-Squared score (0.3915), outperforming MoveNet (0.3545). It also reported the lowest MSE (0.0036) and MAE (0.0473), indicating a more accurate score prediction. However, this interestingly shows a slight discrepancy with the landmark estimation performance where CTML loss is most effective. These results serve as a promising proof-of-concept for automated synchronization scoring using HPE.

V. CONCLUSIONS

In this study, we proposed a synchronized sports scoring framework demonstrated on Taekwondo Poomsae. We built upon the ViTPose architecture and validated the framework using a manually annotated custom dataset. We fine-tuned three ViTPose variants on both the Taekwondo dataset and the LSP benchmark, exploring the impact of seven different loss functions on keypoint detection accuracy. Evaluation using COCO metrics demonstrated that alternatives to the standard MSE loss, particularly CTML and MAE achieved superior mAP and mAR scores. Using Euclidean distances between keypoints, we constructed feature vectors for synchronization scoring and trained regression models to predict expert-

annotated scores. Results showed that ViTPose-based models outperformed MoveNet in synchronization prediction accuracy. Overall, this work establishes a proof-of-concept for synchronized performance evaluation in sports for future advancements in sports analytics and judging automation.

References

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[2] C. Zheng, W. Wu, C. Chen, *et al.*, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[3] A. Fathalla, A. Salah, M. Bekhit, *et al.*, "Real-time and automatic system for performance evaluation of karate skills using motion capture sensors and continuous wavelet transform," *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 1 561 942, 2023.

[4] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 374–382.

[5] M. Latyshev, G. Lopatenko, V. Shandryhos, O. Yarmoliuk, M. Pryimak, and I. Kvasnytsia, "Computer vision technologies for human pose estimation in exercise: Accuracy and practicality," in *SOCIETY. INTEGRATION. EDUCATION. Proceedings of the International Scientific Conference*, vol. 2, 2024, pp. 626–636.

[6] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 571–38 584, 2022.

[7] R. Votel and N. Li, *Next-generation pose detection with movenet and tensorflow.js*, Mar. 2021. [Online]. Available: https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html.

[8] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2014.

[9] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[10] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 977–11 986.

[11] B. Artacho and A. Savakis, *Omnipose: A multi-scale framework for multi-person pose estimation*, 2021. arXiv: 2103.10180 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.10180.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[13] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.

[14] N. T. Thành and P. Công, "An evaluation of pose estimation in video of traditional martial arts presentation," *Journal of Research and Development on Information and Communication Technology*, vol. 2019, no. 2, pp. 114–126, 2019.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[16] W. T. Federation, *World taekwondo poomsae competition rules interpretation*, http://www.worldtaekwondo.org/wp-content/uploads/2019/06/Poomsae_Competition_Rules_and_Interpretation_In_force_as_of_May_14_2019.pdf, Accessed: 2025-07-11, 2019.

[17] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:7318714.

[18] A. Dutta, A. Gupta, and A. Zisserman, *VGG image annotator (VIA)*, "http://www.robots.ox.ac.uk/vgg/software/via/", "Version: 2.0.12, Accessed: 18/1/2025", 2016.

[19] A. Dutta and A. Zisserman, "The via annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2276–2279.

[20] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5700–5709.

[21] S. R. Bakana, Y. Zhang, and B. Twala, "Wildpose: Hrnet-based lightweight and efficient wildlife pose estimation," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2024, pp. 1–6.

[22] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, Springer, 2014, pp. 740–755.