International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Trajectory-Preserving Skeletal Representation for Lightweight Real-Time Gesture Recognition: An Evaluation on SHREC'17

Anurag Sharma
Dept. of Computer Science & Engineering
Shriram College of Engineering & Management
Banmore, Morena (M.P.), India
anuragparashar111@gmail.com

Shyamol Banerjee
Assistant Professor
Dept. of Computer Science & Engineering
Shriram College of Engineering & Management
Banmore, Morena (M.P.), India
shyamolgwl@gmail.com

***Abstract*—Landmark-based gesture recognisers usually translate each hand to its own wrist in every frame, so that the classifier sees hand shape independent of where the hand is. We show that this common step silently removes the information that dynamic gestures depend on. A compact recogniser—a convolutional per-frame encoder, a bidirectional LSTM and a self-attention layer, designed as the hand pathway of a real-time body-and-hand gesture system—reached only 73.8% accuracy on the 14-gesture SHREC'17 benchmark with per-frame wrist anchoring, and per-class analysis showed that the loss fell mainly on swipe and tap gestures. Restoring the wrist trajectory relative to the first frame and adding joint velocities, with every choice made on validation data, raised accuracy to 92.8 ± 0.9% (14 gestures) and 88.6 ± 1.0% (28 gestures) over three seeds, and lifted mean recall on trajectory-defined gestures from 68.9% to 94.1%. With 0.38 M parameters the model lies within the range of published results on this benchmark, above the original SHREC'17 method and Res-TCN, and below the strongest specialised methods by 0.8–3.3 points. Under identical training it is significantly better than an ST-GCN baseline at both granularities and than a single-stream LSTM at 28 gestures. Ablations show that a 30-frame window is sufficient and that the self-attention layer does not improve accuracy on this benchmark. We report the initial and revised results in full, together with the selection procedure, to make the effect of the normalisation choice reproducible.**

***Index Terms*—dynamic hand gesture recognition, skeleton-based recognition, SHREC'17, landmark normalisation, BiLSTM, ablation study, real-time systems**

I. Introduction

Skeleton input has become the default route to gesture recognition on modest hardware. Pose estimators such as MediaPipe [1] turn each video frame into a few dozen landmarks, and a small temporal model on top of them can run alongside the camera on a laptop [2]–[4]. Between the landmarks and the model sits a normalisation step that rarely receives attention. Its usual form is simple: subtract a reference joint (for a hand, the wrist) from every joint in the frame, and optionally divide by a reference length. This removes the dependence on where the user stands and how large the hand appears.

For a static hand pose this is exactly what is wanted. For a dynamic gesture it is not. If the wrist is subtracted in every frame, the wrist sits at the origin throughout the sequence, and a hand that sweeps from left to right becomes indistinguishable from a hand that holds the same shape in one place. Whatever the classifier, the information that separates a swipe from a static pose is gone before the classifier sees the data.

This paper documents that effect on a public benchmark and measures what it costs. The work arose in the development of a real-time recogniser for mixed hand-and-body gestures, whose hand pathway was evaluated on SHREC'17 [5]. The first evaluation, with per-frame wrist anchoring, placed the model fourteen to twenty-one points below published methods. Rather than tune the model, we traced the gap to the representation. The contributions are:

- a per-class analysis on SHREC'17 showing that per-frame wrist anchoring depresses recall specifically on trajectory-defined gestures;
- a revised representation—wrist-relative shape, wrist trajectory relative to the first frame, and joint velocities—selected on validation data, which raises accuracy by about twenty points and places a 0.38 M-parameter recurrent model within the published range;
- a like-for-like comparison with an LSTM and an ST-GCN baseline, with three seeds and paired bootstrap intervals, and an ablation of attention, bidirectionality and window length; and
- a complete account of both evaluation rounds and of how the revision was chosen, so that the result can be checked and reused.

II. Related Work

Skeleton-based recognition. Recurrent models over joint sequences [6] were followed by graph convolutions on the skeleton [7], adaptive graph topologies [8], and channel-wise topology refinement [9], mostly evaluated on body-action datasets such as NTU RGB+D [10]. These models take joint coordinates as given, and the choice of coordinate frame is usually inherited from the dataset's preprocessing.

Dynamic hand gestures. SHREC'17 [5], built from the DHG-14/28 recordings [11], is the standard benchmark for skeleton-based dynamic hand gestures. Its baseline combined

hand-crafted shape and motion descriptors. Res-TCN [12] applied residual temporal convolutions, STA-Res-TCN [13] added spatio-temporal attention, DG-STA [14] learned a dynamic graph with attention over joints and frames, and DD-Net [15] combined joint-collection distances with global motion features at two temporal scales. Several of these methods encode motion explicitly, for instance through frame differences or global trajectory features. Our results suggest why this matters: when motion is not encoded and the coordinate frame removes it, performance on trajectory-defined classes collapses.

Pose-estimator front-ends. MediaPipe Holistic [1], [16] supplies 33 body and 2×21 hand landmarks per frame and has made landmark pipelines practical on CPUs, with known weaknesses in the hand region [17]. Normalising each hand to its wrist is a common choice in such pipelines, which is why the effect studied here is of practical interest beyond SHREC'17.

III. Method

A. System Context

The recogniser studied here is the hand pathway of a real-time system for mixed body-and-hand gestures from a webcam. In the full system, MediaPipe supplies body, left-hand and right-hand landmarks; each stream has its own encoder; the encoded streams are concatenated per frame and passed to a shared temporal backbone; and an auxiliary vector carries anchor positions, visibility and missing-hand flags. SHREC'17 provides a single hand and no body, so only one stream and no auxiliary vector are used here. Fig. 1 shows this single-stream configuration.

B. Input Representations

Let $\boldsymbol{j}_t^{(k)} \in \mathrm{R}^3$ be the 3-D position of hand joint $k \in \{1, \ldots, 22\}$ at frame t, with $k{=}w$ the wrist, and let a gesture be resampled to T frames.

v1 (per-frame wrist anchoring). Each joint is expressed relative to the wrist of the same frame,

$$\boldsymbol{s}_t^{(k)} = \boldsymbol{j}_t^{(k)} - \boldsymbol{j}_t^{(w)}, \tag{1}$$

and the frame vector is $\boldsymbol{x}_t = [\boldsymbol{s}_t^{(1)}; \ldots; \boldsymbol{s}_t^{(22)}] \in \mathrm{R}^{66}$. Because $\boldsymbol{s}_t^{(w)} = \boldsymbol{0}$ for every t, any rigid translation of the hand over time leaves the input unchanged.

v2 (trajectory-preserving). The shape block of (1) is kept and two blocks are added: the wrist position relative to the first frame,

$$\boldsymbol{r}_t = \boldsymbol{j}_t^{(w)} - \boldsymbol{j}_1^{(w)} \in \mathrm{R}^3, \tag{2}$$

and the first temporal difference of every joint,

$$\boldsymbol{d}_t^{(k)} = \boldsymbol{j}_t^{(k)} - \boldsymbol{j}_{t-1}^{(k)}, \qquad \boldsymbol{d}_1^{(k)} = \boldsymbol{0}, \tag{3}$$

giving $\boldsymbol{x}_t = [\boldsymbol{s}_t; \boldsymbol{r}_t; \boldsymbol{d}_t] \in \mathrm{R}^{135}$. The trajectory term retains where the hand goes while remaining invariant to where the gesture starts; the velocity term describes both global motion and finger motion. Neither term requires a body skeleton, so the representation applies to any single-hand landmark source.

TABLE I
Training configuration (identical for all models).

Setting	Value
Window length T	30 frames (resampled)
Optimiser	Adam, lr 10^{-3}, weight decay 10^{-4}
Schedule	cosine annealing, max. 80 epochs
Batch size / gradient clip	64 / ℓ_2 norm 5
Dropout / label smoothing	0.3 / 0.1
Early stopping	val. macro-F_1; patience 10 (v1), 20 (v2)
Seeds	0, 1, 2
Hardware	MacBook Pro M5 Pro, PyTorch (MPS)

C. Recogniser

A two-layer 1-D convolutional encoder with batch normalisation [18], ReLU and a residual connection maps each $\boldsymbol{x}_t$ to a 64-dimensional embedding. A bidirectional LSTM [19] with 128 units per direction processes the T embeddings. A single scaled dot-product self-attention layer [20] with d_k=64 re-weights the hidden states, which are then averaged over time. A fully connected layer of 128 units with ReLU and dropout [21] (p=0.3) and a linear softmax output complete the model. With the v2 input the model has 379,342 trainable parameters.

D. Training

The loss is cross-entropy with label smoothing 0.1. Optimisation uses Adam [22] with learning rate 10^{-3}, weight decay 10^{-4}, cosine annealing over at most 80 epochs, gradient clipping at ℓ_2 norm 5 and batch size 64. Augmentation is applied to the landmarks during training only: temporal rate perturbation within ±20%, additive Gaussian noise, and horizontal reflection. Training stops when validation macro-F_1 has not improved for a fixed number of epochs (the patience), and the best validation checkpoint is restored. Every configuration is trained with seeds 0, 1 and 2. Table I summarises the settings.

IV. Experimental Setup

A. Dataset

SHREC'17 [5] contains 2,800 sequences of 14 dynamic hand gestures performed by 28 participants, each gesture executed with one finger and with the whole hand, which yields a 28-class labelling. A depth camera provides 22 hand joints per frame, and we use their 3-D world coordinates. The official split of 1,960 training and 840 test sequences is used unchanged. About 10% of the training sequences form a validation set; the test set is never used for model selection. The skeleton files were obtained from the public release of the DD-Net code [15], because the official download page was unreachable, and were checked against the official split files: sequence order, labels and per-sequence frame counts match, and the coordinates are the raw, unnormalised world positions. Sequences are resampled to T=30 frames.

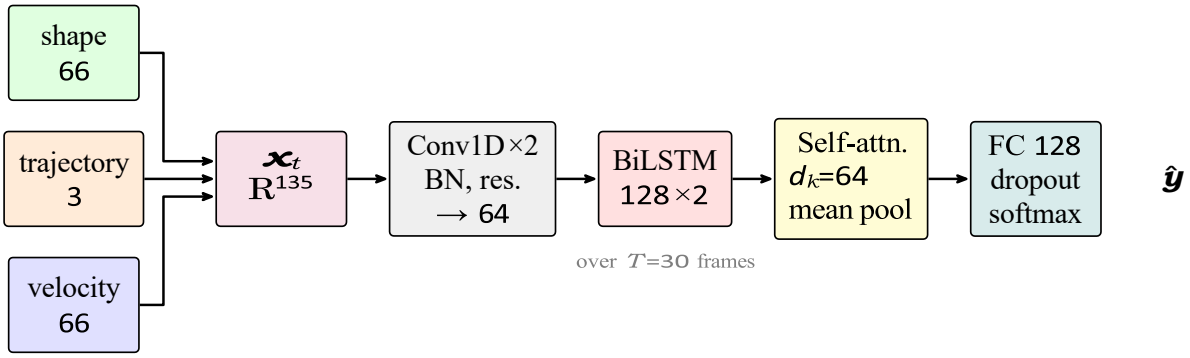


Fig. 1. Single-stream recogniser used on SHREC'17 with the revised (v2) input. The initial (v1) input consisted of the shape block only.

B. Baselines and Ablations

Two baselines share the data pipeline, splits, augmentation and training schedule. **B1** is a single-layer LSTM with 256 units over the same per-frame vector, followed by a linear classifier. **B2** is a compact ST-GCN [7] on the 22-joint hand graph with anatomical edges. Graph node features cannot directly carry the flat 135-dimensional vector, so B2 receives joint coordinates anchored to the first frame of the sequence, which preserves trajectory (3 channels), and a second variant, **B2-vel**, additionally receives per-joint velocity (6 channels). Four ablations change one element of the proposed model: **A1** removes self-attention (the LSTM states are averaged directly), **A3** uses a unidirectional LSTM, and **A5**/**A6** use windows of T=15 and T=60.

C. Metrics and Statistics

We report top-1 accuracy and macro-F_1 on the test set as mean ± sample standard deviation over the three seeds. Differences between the proposed model and each comparator are tested with a paired bootstrap [23] over the 840 test sequences (10,000 resamples) using the seed-0 models; a 95% interval that excludes zero is reported as significant.

D. Procedure

The evaluation proceeded in two rounds. In round one, all models were trained on the v1 input with patience 10 and evaluated on the test set. After the per-class analysis of Section V-B, five candidate revisions were compared on the validation set only (Table III); the best was frozen, and then the proposed model, the baselines and the ablations were trained on it and evaluated once on the test set. We note that the decision to revise the representation was prompted by round-one test results; the choice among revisions used validation data only.

E. Implementation

All experiments ran on a single laptop (Apple MacBook Pro, M5 Pro, 24 GB unified memory) using PyTorch [24] with the Metal (MPS) backend.

TABLE II
ROUND ONE (V1 INPUT): SHREC'17 TEST ACCURACY AND MACRO-F_1 (%), MEAN ± STD OVER THREE SEEDS.

	14 gestures		28 gestures	
Model	Acc.	F_1	Acc.	F_1
B1 LSTM	63.10±0.87	63.00±0.87	57.90±0.15	57.99±0.15
B2 ST-GCN	62.86±2.89	63.02±2.92	61.07±2.05	60.46±1.90
Proposed	73.77±0.20	73.88±0.14	68.85±2.94	68.69±3.08

V. RESULTS

A. Round One: Per-Frame Wrist Anchoring

Table II reports the round-one results. The proposed model was about ten points ahead of both baselines, but all three were far below published results on the same split (Table VII). A deficit of this size shared by three different architectures suggested that the input, rather than any one model, was at fault.

B. Diagnosis

We divided the 14 gestures by what defines them: shape (grab, expand, pinch, rotation clockwise and counter-clockwise), trajectory (tap and the seven swipes), and oscillation (shake, excluded from both groups). If the v1 input erases trajectory, recall on the second group should be lower. In round one, mean recall was 68.9% on trajectory-defined gestures and 80.1% on shape-defined ones. A validation-only check that restored the trajectory while changing nothing else raised validation accuracy by 15–16 points.

C. Selecting the Revision

Table III lists the five candidates, each adding one change to the previous best, compared on validation macro-F_1 averaged over both granularities. The trajectory term alone (V1) recovered most of the deficit, and velocities added two points (V2). Joint-collection distances, as used by DD-Net, reduced validation performance in our model (V3), and a longer window did not help (V4). Raising the patience to 20 gave the best result (V5), which was frozen.

TABLE III
VALIDATION SWEEP (TEST SET UNUSED): MEAN VALIDATION MACRO-F_1 (%) OVER THREE SEEDS, AVERAGED OVER 14 AND 28 GESTURES.

ID	Change	Val. F_1
V1	shape + wrist trajectory	89.99
V2	V1 + joint velocities	92.02
V3	V2 + joint-collection distances	90.74
V4	best + window T=60	90.70
V5	V2 + patience 20	**92.99**

D. Round Two: Test Results

Table IV reports the round-two test results. The proposed model reaches 92.82% at 14 gestures and 88.57% at 28 gestures, gains of 19.05 and 19.72 points over round one. The LSTM baseline gains about 27 points at both granularities, which confirms that the round-one gap was mainly a property of the input.

Fig. 2 shows per-class recall of the proposed model before and after the revision. Mean recall on trajectory-defined gestures rose by 25.2 points to 94.1%, and every swipe now exceeds 91%; shape-defined gestures gained a smaller 9.7 points, to 89.8%, plausibly through the velocity term, which also captures finger motion. Tap remains the weakest class (79.2%), and pinch the weakest shape class (81.0%).

E. Statistical Comparison and Ablations

Table V gives the paired bootstrap intervals. The proposed model is significantly better than both ST-GCN variants at both granularities and than the LSTM at 28 gestures (+4.54 points). At 14 gestures its three-seed mean exceeds the LSTM by 2.38 points, but the seed-0 interval includes zero, so this difference is not established. The ST-GCN results at 28 gestures should not be read as a property of graph models: B2 and B2-vel did not receive the full representation, and the 28-class task, which separates finger from whole-hand executions, appears to need it.

None of the ablations (Table VI) differs significantly from the full model. Removing attention (A1) does not reduce accuracy and slightly raises it at 28 gestures while saving about 99 k parameters; on this single-stream benchmark the attention layer does not earn its cost. A unidirectional LSTM (A3) is 3.1 points lower on the three-seed mean at 14 gestures, with an interval that only just includes zero, and equal at 28 gestures. Halving the window (A5) changes nothing measurable, and doubling it (A6) is one to two points lower. For a real-time system, where a shorter window means lower latency, this favours T=15 to 30.

F. Comparison with Published Methods

Table VII compares the frozen model with results reported on the same split. Published values are quoted from the original papers. The proposed model is above the original SHREC'17 method [5] by 4.6 and 6.7 points and above Res-TCN [12] by 1.7 and 1.3 points. It is below STA-Res-TCN, DG-STA and DD-Net by 0.8–1.8 points at 14 gestures and 2.1–3.3 points at 28 gestures. Those methods are designed for depth-sensor hand skeletons; the present model is the hand pathway of a webcam system for body and hands, evaluated with a fixed, mostly untuned recipe. Since published figures are frequently single runs, differences of one to two points against a three-seed mean should be read with care.

VI. DISCUSSION

Normalisation as a modelling decision. The largest effect in this study came from neither the architecture nor the training recipe but from a preprocessing step, and it affected all three model families. Per-frame anchoring to a reference joint is appropriate for static poses and for streams whose position is carried elsewhere—in our full system the anchor coordinates travel in a separate auxiliary vector. It is harmful whenever it is the only representation of a moving part. The remedy is cheap: three trajectory values and one difference per joint coordinate. We expect the same issue to arise in other single-hand landmark pipelines, including those built on MediaPipe hand landmarks, and suggest that per-class recall on trajectory-defined gestures is a quick test for it.

What the architecture contributes. With a good input, the margins between models shrink. The proposed model is clearly better than our ST-GCN baselines, but the latter were not given the full representation, and its advantage over a plain LSTM is significant only on the finer 28-class task. The ablations indicate that bidirectionality is the most useful element of the temporal backbone, while self-attention adds parameters without measurable benefit here. Whether attention helps when it must weigh several streams, as in the full body-and-hand system, is an open question for our mixed-gesture dataset.

Limitations. The revision was prompted by test-set evidence from round one, although it was selected on validation data; readers should weigh the final numbers accordingly, and we report round one in full for that reason. Bootstrap intervals use the seed-0 models only, and three seeds characterise variability only coarsely. The ST-GCN baselines received a reduced form of the representation. SHREC'17 exercises one hand pathway on depth-sensor skeletons; results on webcam landmarks, on the full three-stream model, and end-to-end latency are not reported here and are the subject of ongoing work.

VII. CONCLUSION

Per-frame wrist anchoring, a common step in landmark-based gesture recognition, erases the hand trajectory and with it most of the signal that separates dynamic gestures. On SHREC'17 this cost a compact BiLSTM recogniser about twenty accuracy points and twenty-five points of recall on trajectory-defined gestures. Restoring the trajectory relative to the first frame and adding joint velocities, chosen on validation data, brought a 0.38 M-parameter model to 92.8% and 88.6% on the 14- and 28-gesture tasks, within the range of specialised published methods. Ablations showed that a short window suffices and that self-attention did not help on this benchmark. Future work will evaluate the full body-and-hand model on a

TABLE IV
ROUND TWO (V2 INPUT, FROZEN CONFIGURATION): SHREC'17 TEST ACCURACY AND MACRO-F_1 (%), MEAN ± STD OVER THREE SEEDS, AND TRAINABLE PARAMETERS.

	14 gestures		28 gestures		
Model	Acc.	F_1	Acc.	F_1	Params
B1 LSTM	90.44 ± 0.48	90.26 ± 0.45	84.88 ± 0.83	84.58 ± 0.83	406,030
B2 ST-GCN	90.75 ± 1.38	90.51 ± 1.39	71.15 ± 0.79	70.30 ± 0.85	217,134
B2-vel	90.00 ± 0.68	89.79 ± 0.66	72.34 ± 0.62	71.56 ± 0.79	217,326
Proposed	**92.82 ± 0.92**	**92.63 ± 0.98**	**88.57 ± 0.99**	**88.43 ± 1.02**	379,342

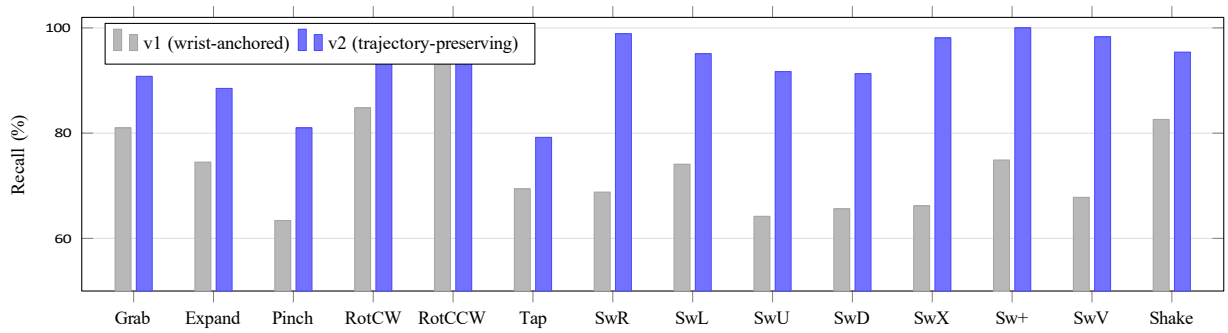


Fig. 2. Per-class test recall of the proposed model at 14 gestures, before and after the revision, over the predictions of all three seeds. The first five classes are shape-defined, the next eight trajectory-defined.

TABLE V
PAIRED BOOTSTRAP ON THE TEST SET: ACCURACY OF THE PROPOSED MODEL MINUS THE COMPARATOR (POINTS), SEED-0 MODELS, 95% CI. *CI EXCLUDES ZERO.

	14 gestures		28 gestures	
Comparator	Diff.	95% CI	Diff.	95% CI
B1 LSTM	+0.97	[−1.07, 3.10]	+4.54*	[1.90, 7.14]
B2 ST-GCN	+2.87*	[0.60, 5.12]	+18.19*	[14.88, 21.55]
B2-vel	+2.28*	[0.12, 4.40]	+16.40*	[13.21, 19.64]
A1 no attn.	−1.30	[−3.10, 0.48]	−0.82	[−2.98, 1.31]
A3 unidir.	+1.81	[−0.12, 3.81]	0.00	[−2.14, 2.02]
A5 T=15	0.00	[−1.67, 1.67]	−0.48	[−2.50, 1.55]
A6 T=60	+0.85	[−1.07, 2.74]	+1.78	[−0.24, 3.69]

TABLE VI
ABLATIONS (V2 INPUT): TEST ACCURACY (%), MEAN ± STD OVER THREE SEEDS.

Variant	14 gestures	28 gestures	Params
Full model	92.82±0.92	88.57±0.99	379 k
A1 no attention	92.58±0.56	89.64±0.45	281 k
A3 unidirectional	89.68±0.46	89.09±0.98	198 k
A5 T=15	92.58±0.66	89.56±0.63	379 k
A6 T=60	90.67±0.22	87.58±0.73	379 k

subject-disjoint webcam dataset, measure end-to-end latency on a laptop, and test whether attention becomes useful when several landmark streams must be fused.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science & Engineering, Shriram College of Engineering & Management, for its support.

TABLE VII
ACCURACY (%) ON SHREC'17. PUBLISHED VALUES FROM THE ORIGINAL PAPERS; LAST ROW MEASURED HERE (MEAN OF THREE SEEDS).

Method	14 gestures	28 gestures
De Smedt et al. [5]	88.2	81.9
Res-TCN [12]	91.1	87.3
STA-Res-TCN [13]	93.6	90.7
DG-STA [14]	94.4	90.7
DD-Net [15]	94.6	91.9
Ours, v1 input	73.8	68.9
Ours, v2 input	92.8	88.6

DECLARATION ON THE USE OF AI TOOLS

A large language model-based assistant (Claude, Anthropic) was used to assist with drafting and editing portions of the text, with LaTeX formatting, and with writing experiment code. All experiments were run by the authors, all reported numbers are taken from the experiment logs, and the authors take full responsibility for the manuscript.

REFERENCES

[1] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
[2] B. Sundar and T. Bagyammal, "American Sign Language recognition for alphabets using MediaPipe and LSTM," in *Procedia Computer Science*, vol. 215, 2022, pp. 642–651.
[3] S. Kamble, "SLRNet: A real-time LSTM-based sign language recognition system," *arXiv preprint arXiv:2506.11154*, 2025.
[4] M. Garg *et al.*, "MP-GestLSTM: Real-time gesture detection using MediaPipe and LSTM," *Systems Science & Control Engineering*, 2025.

[5] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "SHREC'17 track: 3D hand gesture recognition using a depth and skeletal dataset," in *Eurographics Workshop on 3D Object Retrieval*, 2017.
[6] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 816–833.
[7] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, vol. 32, 2018, pp. 7444–7451.
[8] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 026–12 035.
[9] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *IEEE/CVF International Conference on Computer Vi-sion (ICCV)*, 2021, pp. 13 359–13 368.
[10] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019.
[11] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 1206–1214.
[12] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1623–1631.
[13] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention Res-TCN for skeleton-based dynamic hand gesture recognition," in *European Conference on Computer Vision Workshops (ECCVW)*, 2018.
[14] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, "Construct dynamic graphs for hand gesture recognition via spatial-temporal attention," in *British Machine Vision Conference (BMVC)*, 2019.
[15] F. Yang, S. Sakti, Y. Wu, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *ACM Multimedia Asia*, 2019.
[16] 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.
[17] A. Moryossef *et al.*, "Optimizing hand region detection in MediaPipe holistic full-body pose estimation to improve accuracy and avoid downstream errors," *arXiv preprint arXiv:2405.03545*, 2024.
[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.
[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[23] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1994.
[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.