OO-dMVMT: A Deep Multi-view Multi-task Classification Framework for Real-time 3D Hand Gesture Classification and Segmentation
-*Supplementary Material*-

Federico Cunico[*] Federico Girella[*] Andrea Avogaro[*]
Marco Emporio[*] Andrea Giachetti Marco Cristani

University of Verona

name.surname@univr.it

The supplementary material is structured as follows:

- **Qualitative results** (Sec. 1): a short video is available as additional material (OO_dMVMT.mp4, refer to Sec. 1 for the download link). This section represents an in-depth look at what is shown in the video, and can be read before or after watching it. Also, a short video of the alpha version of the Hololens2 real-time demo is available at the same link. The demo will be further improved in the user experience and released upon acceptance;
- **Extensive results on SHREC'22** (Sec. 2): we present additional results on the SHREC'22 benchmark. Specifically, disaggregated results (per gesture class) in terms of Detection Rate (DR), False Positives (FP) and Jaccard Index (JI) will be presented in Tab. 1, Tab. 2 and Tab. 3, respectively. In Sec. 2.1 we explore the performance in terms of False Positives. In Sec. 2.2 we will give supplementary results w.r.t. Minimum Overlap Ratio. Finally, in Sec. 2.3 we investigate the impact of different values for W, as an additional ablation study.
- **Extensive results on SHREC'19** (Sec. 3): some additional results on SHREC'19 will be presented. In particular, Tab. 4 and Tab. 5 show the Detection Rate and False Positive scores per class, respectively. In Sec. 3.1 we study the results in terms of Detection Rate per class on SHREC'19. In Sec. 3.2 we present the false positives for each model, again divided by class.

1. Qualitative results

At the link https://tinyurl.com/oomvmt2023, two videos are available. In the video file OO_dMVMT.mp4 we primarily want to show how the online scenario for 3D gesture classification works in practice, and how our method performs. In a genuine human-computer interaction session, gestures are fast, heterogeneous, and interleaved with noisy natural movements (namely *non-gestures*). The SHREC'22 benchmark mirrors perfectly these conditions: it is largely composed of non-gestures, consisting of random poses or/and trajectories that may resemble in some cases gestures contained in the dictionary. This makes the classification task definitely hard. In the sequences in which only OO-dMVMT is reported, the idea is to communicate the responsiveness of our approach. In the other sequences, we show our approach with another comparative method, to highlight specific advantages of OO-dMVMT w.r.t. the state of the art.

The video file RealtimeDemo_v0.1__LAB.mp4 shows a short video of the first version of the online demo. In the middle of the video, it is possible to notice that there is still some uncertainty between the GRAB gesture and PINCH gesture in the initial movements. In this video, the demo software was under development, and an improved version in terms of user experience and response time will be developed.

2. Extensive results on SHREC'22

In this section, we present additional results on the SHREC'22 [4] dataset. For a tabular view of the data, the per-gesture Detection Rates are in Tab. 1, the False Positive Rates per gesture are shown in Tab. 2, while Tab. 3 reports the Jaccard Index results per gesture class. Also, the confusion matrix of our method and the best-performing comparative method DDNet [10] are shown in Fig. 1 and Fig. 2. As expected, dynamic gestures, characterized by long trajectories sharing similar sub-parts can be confused with each other, and fine-grained dynamic gestures with similar features like PINCH and GRAB are sometimes interchanged.

[*] The authors contributed equally to this paper

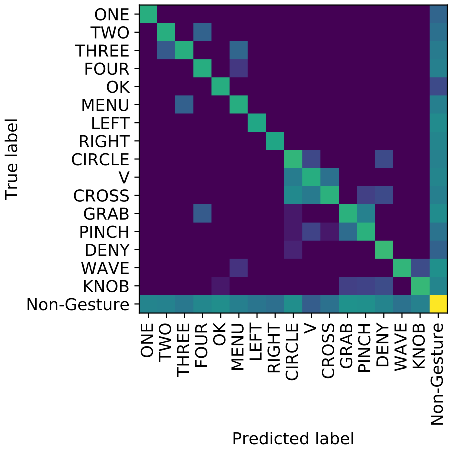

Figure 1. Confusion matrix (log scale color map) of OO-dMVMT on SHREC'22

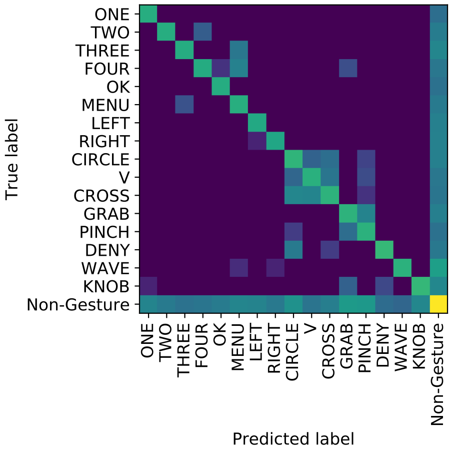

Figure 2. Confusion matrix (log scale color map) of DDNet [10] on SHREC'22

The fact that the labels of a subset of static gestures (TWO, THREE, FOUR, MENU) are also sometimes switched between them suggests that the descriptors of hand pose can still be improved. In any case, our method results in a lower number of wrong labels.

2.1. False Positive Rate

In Tab. 2 the False Positive Rates for each model are shown (lower is better). When paired with the detection rates in Tab. 1, this data shows how, even though we are achieving state-of-the-art results globally, specific actions have other methods as the best-performing ones.

In particular, in detection rate we have a big advantage w.r.t. the other approaches with the DENY, PINCH, and CROSS gestures. On the other side, for False Positive Rate, we tend to have good scores for all the gestures, even though we don't have the absolute best in some classes such as CIRCLE, PINCH, and KNOB.

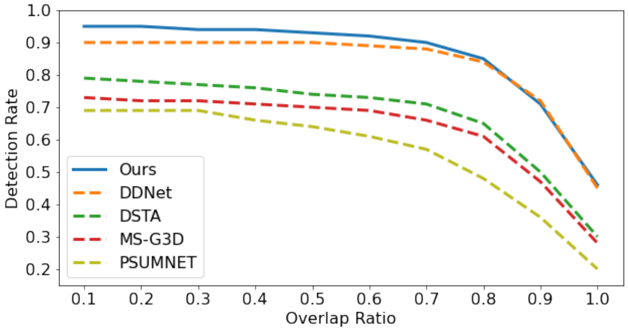

Figure 3. Detection Rate (↑) as a function of MOR on SHREC'22. Higher is better. Only the five best methods for DR (with code available) are shown for clarity

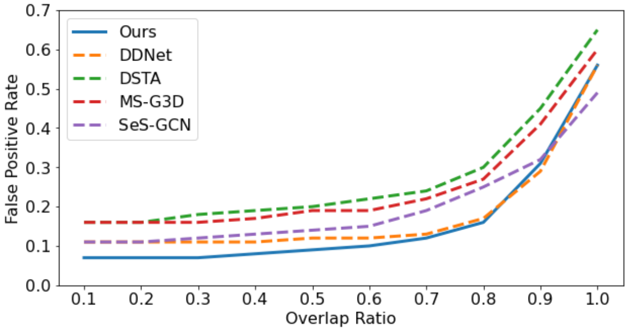

Figure 4. False Positive Rate (↓) as a function of MOR on SHREC'22. Lower is better. Only the five best methods for FP (with code available) are shown for clarity

2.2. Minimum Overlap Ratio

Fig. 3 and Fig. 4 show, respectively, the DR and FP as a function of the Minimum Overlap Ratio (MOR). As explained in the main paper, the MOR parameter ranges from 0 to 1, where 1 enforces a complete overlap (on the time dimension) of the detected gesture with the Ground Truth one, while 0 completely removes this temporal constraint. The results shown in these plots are consistent with the one shown in the main paper (Jaccard Index as function of the MOR, Fig. 4), showing an interesting fact: whereas the task is easier (i.e. setting a low overlap ratio) we are significantly better than all the competitors in both the DR and FP scores. When MOR increases, we become similar to the behavior of the best-considered comparative methods.

2.3. The impact of W

As an additional ablation study on SHREC'22, we tested OO-dMVMT using different values for W, thus changing

Table 1. Detection Rate (↑) per gesture on SHREC'22. Higher is better. We underline the cases in which OO-dMVMT is the second best result.

Model	**ONE**	**TWO**	**THREE**	**FOUR**	**OK**	**MENU**	**LEFT**	**RIGHT**	**CIRCLE**	**V**	**CROSS**	**GRAB**	**PINCH**	**DENY**	**WAVE**	**KNOB**	**AVG**
2ST-GCN 5F [4]	0.92	0.83	0.75	0.86	0.69	0.72	0.39	0.31	0.89	0.58	0.69	0.78	0.81	0.89	**0.83**	0.83	0.74
Causal TCN [4]	0.92	0.83	0.75	**0.89**	0.69	0.72	0.39	0.31	0.89	0.58	0.69	0.78	0.81	0.89	**0.83**	0.83	0.80
TN-FSM+JD [4]	0.97	0.78	0.56	0.47	0.56	0.89	0.69	0.72	0.92	0.89	0.81	**0.86**	0.78	0.86	0.69	0.89	0.77
Stronger [4]	0.92	0.72	0.86	0.75	0.75	**0.97**	0.67	0.94	0.72	**0.94**	0.03	0.64	0.86	0.78	0.36	0.58	0.72
DG-STA [3]	0.67	0.61	0.56	0.56	0.47	0.33	0.56	0.47	0.31	0.17	0.72	0.50	0.33	0.78	0.44	0.72	0.51
SeS-GCN [7]	0.75	0.61	0.56	0.64	0.50	0.86	0.67	0.53	0.36	0.58	0.69	0.50	0.39	0.81	0.72	0.72	0.60
DDNet [10]	**1.00**	**0.97**	0.92	0.72	**1.00**	0.94	**1.00**	**0.97**	**0.94**	0.86	0.81	0.72	0.86	0.89	0.78	0.97	0.88
MS-G3D [5]	0.78	0.69	0.64	0.64	0.64	0.86	0.64	0.78	0.69	0.72	0.64	0.39	0.67	0.78	0.72	0.86	0.68
PSUMNET [9]	0.61	0.94	0.69	0.61	0.64	0.47	0.67	0.75	0.75	0.50	0.58	0.39	0.44	0.81	0.67	0.72	0.62
DeepGru [6]	0.28	0.31	0.28	0.14	0.08	0.19	0.56	0.33	0.44	0.56	0.39	0.22	0.06	0.22	0.19	0.31	0.26
DSTA [8]	0.94	0.83	0.81	0.64	0.78	0.83	0.75	0.72	0.61	0.53	0.81	0.58	0.75	0.83	0.75	0.75	0.73
OO-dMVMT	**1.00**	**0.97**	**0.97**	0.86	**1.00**	0.92	0.97	**0.97**	0.89	0.83	**0.89**	0.81	**0.92**	**0.97**	0.70	**1.00**	**0.92**

Table 2. False Positive Rate (↓) per gesture on SHREC'22. Lower is better. We underline the cases in which OO-dMVMT is the second best result.

Model	**ONE**	**TWO**	**THREE**	**FOUR**	**OK**	**MENU**	**LEFT**	**RIGHT**	**CIRCLE**	**V**	**CROSS**	**GRAB**	**PINCH**	**DENY**	**WAVE**	**KNOB**	**AVG**
2ST-GCN 5F [4]	0.22	0.08	0.19	0.14	0.22	0.25	0.47	0.47	**0.08**	0.31	0.19	0.19	0.28	0.17	0.28	0.19	0.23
Causal TCN [4]	**0.00**	**0.00**	**0.00**	0.06	**0.00**	**0.00**	0.19	0.31	0.47	0.11	1.25	0.22	0.11	0.03	0.11	1.22	0.29
TN-FSM+JD [4]	0.08	**0.00**	**0.00**	0.06	0.03	0.08	0.25	0.53	0.44	0.42	0.06	0.28	0.19	0.06	0.06	0.39	0.23
Stronger [4]	0.03	**0.00**	**0.00**	0.06	0.08	0.03	1.64	0.17	0.75	0.47	**0.00**	0.36	0.14	0.06	0.61	0.89	0.34
DG-STA [3]	0.06	0.19	0.17	0.03	0.06	0.03	0.19	0.17	0.42	0.44	1.64	0.31	0.25	0.33	0.06	0.64	0.32
SeS-GCN [7]	0.06	0.03	**0.00**	0.06	0.03	0.03	0.14	0.14	0.14	0.25	0.39	0.31	0.11	0.11	0.03	0.44	0.16
DDNet [10]	0.03	**0.00**	0.03	0.06	0.03	0.19	0.06	0.06	0.25	0.14	0.11	0.56	0.28	**0.00**	0.08	**0.00**	0.16
MS-G3D [5]	**0.00**	**0.00**	**0.00**	0.08	0.03	0.06	0.08	0.36	0.31	0.69	0.28	0.25	0.40	0.19	**0.00**	0.25	0.21
PSUMNET [9]	0.06	0.03	0.11	0.03	0.03	**0.00**	0.06	0.11	0.56	0.75	0.28	0.19	0.50	0.03	0.03	0.56	0.24
DeepGru [6]	**0.00**	0.06	0.14	**0.00**	**0.00**	0.03	**0.03**	0.06	0.58	1.83	0.31	0.19	**0.06**	0.14	0.33	0.42	0.25
DSTA [8]	0.03	0.03	0.11	0.06	0.06	0.03	0.14	0.11	0.50	0.36	0.44	**0.14**	0.53	0.03	0.22	0.50	0.24
OO-dMVMT	**0.00**	**0.00**	0.02	0.05	0.02	0.02	0.11	**0.02**	0.25	**0.00**	0.19	**0.14**	0.33	**0.00**	0.02	0.33	**0.09**

the number of frames that our model had access to while making its decision. The study has been conducted by splitting the SHREC'22 train set into a train and validation set, by uniform sampling with the following proportions: 80% train, 20% validation. We show the results in Fig. 5, Fig. 6 and Fig. 7, where curves are presented for Detection Rate, False Positives and Jaccard Index respectively. We can notice how, with $W = 30$, OO-dMVMT performs better than the reported results in the main paper, which have been obtained with $W = 16$. However, due to the post-processing step applied to the results (see Sec.3 of main paper), a longer window would result in a longer delay after which we provide the output of the classification. In other words, $W = 16$ has been individuated as an optimal compromise to provide good performance on all the possible figures of merit.

A small W sharply reduces all the scores. This is naturally expected, since longer gestures like CIRCLE, DENY and others require longer windows.

3. Extensive results on SHREC'19

In this section, we deepen the analysis on the SHREC'19 [1] benchmark, with results left out of the main paper for the sake of space.

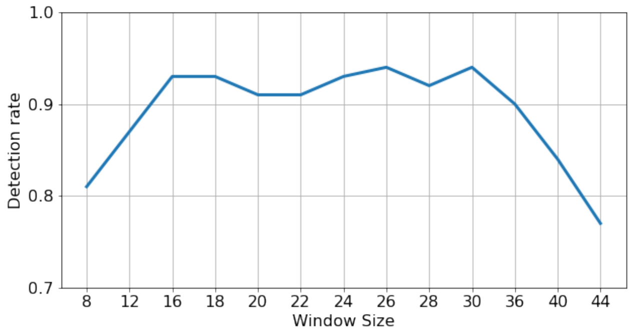

Figure 5. Detection Rate (↑) of OO-dMVMT as a function of the window size W on SHREC'22. Higher is better.

3.1. Detection Rate

In Fig. 8 and Tab. 4 we show the detection rates per gesture obtained with all the classification models with publicly available code, trained with the same protocol followed in the main paper.

While OO-dMVMT is suboptimal in some gestures

Table 3. Jaccard Index (↑) per gesture on SHREC'22. Higher is better. We underline the cases in which OO-dMVMT is the second best result.

Model	ONE	TWO	THREE	FOUR	OK	MENU	LEFT	RIGHT	CIRCLE	V	CROSS	GRAB	PINCH	DENY	WAVE	KNOB	AVG
2ST-GCN 5F [4]	0.75	0.77	0.63	0.76	0.57	0.58	0.26	0.21	**0.82**	0.45	0.58	0.65	0.63	0.76	0.67	0.70	0.61
Causal TCN [4]	0.97	**0.97**	**0.94**	**0.89**	**0.97**	0.94	0.58	0.40	0.31	0.21	0.39	0.62	**0.83**	0.78	0.55	0.44	0.68
TN-FSM+JD [4]	0.88	0.70	0.50	0.40	0.48	0.82	0.53	0.46	0.63	0.63	**0.76**	0.65	0.64	0.82	0.63	0.60	0.63
Stronger [4]	0.89	0.72	0.86	0.71	0.69	**0.95**	0.25	0.81	0.40	0.64	0.03	0.45	0.76	0.74	0.22	0.31	0.59
DG-STA [3]	0.63	0.51	0.48	0.54	0.45	0.32	0.47	0.40	0.22	0.12	0.27	0.38	0.27	0.58	0.42	0.44	0.40
SeS-GCN [7]	0.71	0.59	0.56	0.61	0.49	0.84	0.59	0.46	0.32	0.47	0.50	0.38	0.35	0.73	0.70	0.50	0.53
DDNet [10]	0.97	**0.97**	0.89	0.68	**0.97**	0.79	**0.95**	0.92	0.76	0.76	0.73	0.46	0.67	0.89	**0.72**	**0.97**	0.78
MS-G3D [5]	0.78	0.69	0.64	0.59	0.62	0.82	0.59	0.57	0.53	0.43	0.50	0.31	0.48	0.65	**0.72**	0.69	0.57
PSUMNET [9]	0.58	0.92	0.63	0.59	0.62	0.47	0.63	0.68	0.48	0.29	0.46	0.33	0.30	0.78	0.65	0.46	0.52
DeepGru [6]	0.28	0.29	0.24	0.14	0.08	0.19	0.54	0.32	0.28	0.20	0.30	0.19	0.05	0.20	0.15	0.22	0.21
DSTA [8]	0.92	0.81	0.73	0.61	0.74	0.81	0.66	0.65	0.41	0.39	0.56	0.51	0.49	0.81	0.61	0.50	0.61
OO-dMVMT	**1.00**	**0.97**	**0.94**	0.81	**0.97**	0.89	0.87	**0.94**	0.71	**0.83**	0.74	**0.71**	0.69	**0.97**	0.67	0.75	**0.85**

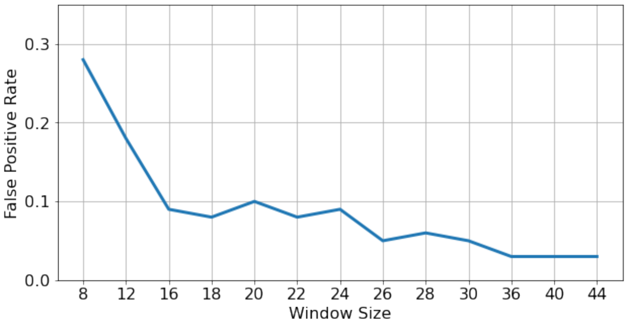

Figure 6. False Positive Rate (↓) of OO-dMVMT as a function of the window size W on SHREC'22. Lower is better.

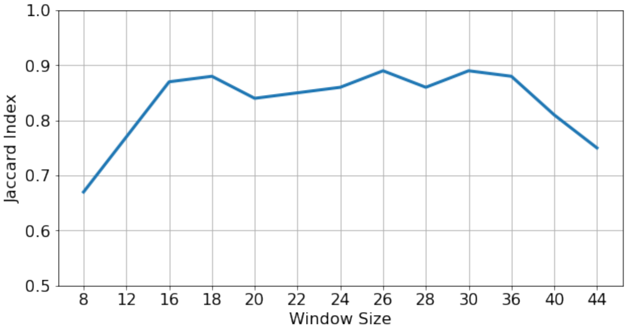

Figure 7. Jaccard Index (↑) of OO-dMVMT as a function of the window size W on SHREC'22. Higher is better.

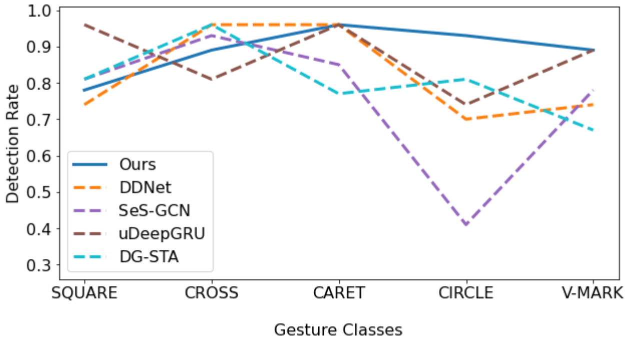

Figure 8. Detection Rate (↑) per class on SHREC'19. Higher is better. Only the best five methods are shown for clarity.

(SQUARE, CARET, and V-MARK), it obtains the best performances on CIRCLE and CROSS. In particular, the latter clearly outperforms all the other methods. This is likely due to the disambiguation of potentially similar trajectory subparts provided by the regression tasks focusing on the detection of gestures' start and end. As shown in the ablation study in Sec. 4.4.1 of the main paper, the regression heads help OO-dMVMT focus on the difference between non-gesture and the start/end of a gesture, resulting in a stronger detection (and classification) performance.

Table 4. Detection Rate (↑) per gesture on SHREC'19. Higher is better. We underline the cases in which OO-dMVMT is the second best result.

Model	SQUARE	CROSS	CARET	CIRCLE	V-MARK	AVG
SW 3-cent [2]	0.74	0.78	0.89	0.48	**0.89**	0.76
uDeepGRU [1]	**0.96**	0.81	**0.96**	0.74	0.78	0.85
DG-STA [3]	0.81	0.96	0.77	0.81	0.67	0.81
SeS-GCN [7]	0.81	0.93	0.85	0.41	0.78	0.75
DDNet [10]	0.74	0.96	**0.96**	0.70	0.74	0.82
MS-G3D [5]	0.74	0.93	0.77	0.30	0.74	0.69
PSUMNET [9]	0.70	0.85	0.69	0.30	0.67	0.64
DSTA [8]	0.81	0.89	0.88	0.63	0.81	0.81
OO-dMVMT	0.81	**1.00**	0.88	**0.82**	0.85	**0.88**

3.2. False Positive Rate

In Fig. 9 and Tab. 5, we show the number of false positives (per class, lower is better) for each model whose code was available. Since the dictionary of SHREC'19 is composed of gestures of a single type (which would be classified as ”dynamic” in SHREC'22), this figure differs from Fig. 5 of the main paper, as the stacked elements here rep-

Table 5. False Positive Rate (↓) per gesture on SHREC'19. Lower is better. Missing values (not reported in the original benchmark [1]) are shown as (–). We underline the cases in which OO-dMVMT is the second best result.

Model	SQUARE	CROSS	CARET	CIRCLE	V-MARK	AVG
SW 3-cent [2]	–	–	–	–	–	0.19
uDeepGRU [1]	–	–	–	–	–	0.10
DG-STA [3]	**0.11**	**0.00**	0.04	**0.07**	0.11	0.07
SeS-GCN [7]	**0.11**	0.04	**0.00**	0.33	0.11	0.12
DDNet [10]	0.22	**0.00**	0.04	0.15	0.07	0.10
MS-G3D [5]	0.19	0.07	0.15	0.63	0.19	0.25
PSUMNET [9]	0.19	0.04	0.15	0.56	0.19	0.22
DSTA [8]	**0.11**	0.04	**0.00**	0.19	0.07	0.08
OO-dMVMT	**0.11**	**0.00**	0.03	**0.07**	**0.03**	**0.05**

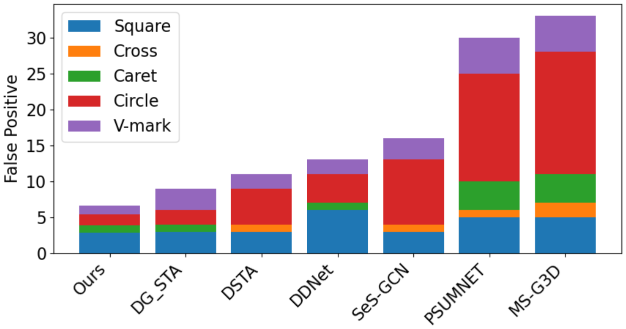

Figure 9. False Positives (↓) per class on SHREC'19. Lower is better.

resent classes instead of gesture types. The performance of OO-dMVMT is far better than the other approaches, especially in the V-MARK class.

References

[1] Fabio Marco Caputo, S Burato, G Pavan, Théo Voillemin, Hazem Wannous, Jean-Philippe Vandeborre, Mehran Maghoumi, EM Taranta, A Razmjoo, JJ LaViola Jr, et al. Shrec 2019 track: online gesture recognition. In *Eurographics Workshop on 3D Object Retrieval*. The Eurographics Association, 2019. 3, 4, 5

[2] Fabio M Caputo, Pietro Prebianca, Alessandro Carcangiu, Lucio D Spano, and Andrea Giachetti. Comparing 3d trajectories for simple mid-air gesture recognition. *Computers & Graphics*, 73:17–25, 2018. 4, 5

[3] Yuxiao Chen, Long Zhao, Xi Peng, Jianbo Yuan, and Dimitris N Metaxas. Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. *arXiv preprint arXiv:1907.08871*, 2019. 3, 4, 5

[4] Marco Emporio, Ariel Caputo, Andrea Giachetti, Marco Cristani, Guido Borghi, Andrea D'Eusanio, Minh-Quan Le, Hai-Dang Nguyen, Minh-Triet Tran, Felix Ambellan, et al. Shrec 2022 track on online detection of heterogeneous gestures. *Computers & Graphics*, 107:241–251, 2022. 1, 3, 4

[5] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 3, 4, 5

[6] Mehran Maghoumi and Joseph J. LaViola Jr. Deepgru: Deep gesture recognition utility. *CoRR*, abs/1810.12514, 2018. 3, 4

[7] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. Pose forecasting in industrial human-robot collaboration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 51–69. Springer, 2022. 3, 4, 5

[8] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3, 4, 5

[9] Neel Trivedi and Ravi Kiran Sarvadevabhatla. Psumnet: Unified modality part streams are all you need for efficient pose-based action recognition. *arXiv preprint arXiv:2208.05775*, 2022. 3, 4, 5

[10] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019. 1, 2, 3, 4, 5