Supplementary Material: Context-Aware Sequence Alignment using 4D Skeletal Augmentation

Taein Kwon[1] Bugra Tekin[2] Siyu Tang[1] Marc Pollefeys[1,2]
[1]Department of Computer Science, ETH Zürich [2]Microsoft MR & AI Lab, Zürich

In the supplemental material, we first provide details about the temporally smoothed noise we used for 4D augmentation and our hyper-parameters. Next, we analyze the performance of our approach for fine-grained frame retrieval and online sequence alignment. We then provide additional qualitative results of our algorithm for aligning two sequences and discuss our design choices for VPoser latent space, phase classification and encoding contextual informaton. Finally, we discuss the broader social impact of our work. Additional qualitative visual results can be found in the accompanying video. Note that, in the accompanying video, we align the sequences by finding nearest neighbors in the embedding space without any post-processing.

S.1. Temporally Smoothed Noise

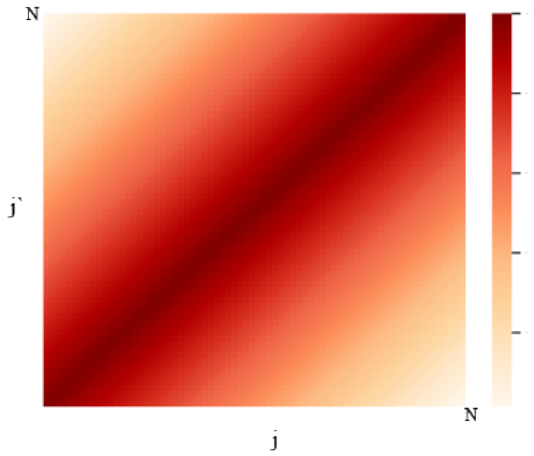

Figure S1. **Covariance matrix for our zero-mean multivariate normal noise distribution.**

While augmenting the joint angle and the latent space, the amount of noise applied across consecutive frames should not be completely independent of each other to preserve the smoothness and consistency of motion. To this end, we propose to add temporally smoothed noise, $MN(C)$, across the sequence, as explained in our paper. We model this using a multivariate normal distribution with a covariance matrix C that enforces high correlation between temporally close frames within the same augmented sequence and low correlation between frames that are further away from each other. We depict the covariance matrix in Fig. S1 and formulate it as follows:

$$C_{j,j'} = 1 - \frac{|j - j'|}{2 \cdot N}, \tag{1}$$

where j and j' depict two frame indices from the augmented sequence, and N is the length of the augmented sequence. When j and j' are close to each other, the covariance is high, indicating that the noise applied on the poses at those frames are similar. This eventually results in less jittery and smooth augmented sequences.

S.2. Implementation Details

We list the hyperparameters we use in our experiments in Table S1.

Hyperparameter	Value
Batch Size	64 (Penn), 32 (H2O), 4 (IKEA)
Learning rate	3e-3(Penn), 3e-4 (H2O), 3e-2 (IKEA)
Optimizer	ADAM
Temperature (λ_{temp})	0.1
3D geometric noise probability	0.3
Noise standard deviation (σ)	10 (angle), 0.1 (VPoser, translation)
Number of attention layers (N_{att})	4
Number of heads (parallel attention layers)	15 (Penn), 17(IKEA), 21 (H2O)
Frames per second	20 (Penn), 30 (IKEA, H2O)

Table S1. **Hyperparameters in our experiment.**

S.3. Fine-Grained Frame Retrieval

We show fine-grained frame retrieval results in Table S2. We find the K nearest frames from one query frame in the embedding space. Following [3], we report Average Precision (AP) at K, that is, the average percentage of correctly retrieved action phase labels within K-retrieved frames. Table S2 shows that our method improves upon prior work by a large margin (an improvement of 10.77% at $K = 5$, 10.46% at $K = 10$ and 10.17% at $K = 15$). In Fig. S3, we show qualitative results of our algorithm compared to TCC [2]. We observe that our method is able to accurately retrieve relevant frames by reasoning about the temporal context of the actions.

Method	AP@5	AP@10	AP@15
SAL [4]	76.04	75.77	75.61
TCN [6]	77.84	77.51	77.28
TCC [2]	76.74	76.27	75.88
LAV [3]	79.13	78.98	78.90
CASA (Ours)	**89.90**	**89.44**	**89.07**

Table S2. **Fine-grained frame retrieval.** We compare fine-grained frame retrieval results on the Penn Action dataset [8].

	Offline	Online	TCC [2]	LAV [3]
Phase classification	92.20	88.01	81.35	84.25
Phase progress	0.9449	0.8454	0.6638	0.6613
Kendall's Tau (τ)	0.9728	0.9059	0.7012	0.8047

Table S3. **Ablation study of online sequence alignment.** We compare the phase classification, phase progress, and Kendall's tau for online and offline operating modes of our model on the Penn Action dataset [8]. While we use the full sequence for offline mode, we only use the embeddings up until the current frame for the online mode.

S.4. Online Sequence Alignment

Our method uses an attention-based model to capture context from all the frames to compute alignment across two videos. However, for online applications, the assumption of having the full sequence will not be valid. Therefore, to demonstrate the potential of our approach for online applications (*e.g.*, online task guidance in augmented reality), we perform an additional experiment, in which, we use contextual information only using frames, seen until the current time frame. To this end, we rely on embeddings computed until the current frame and use it for matching across sequences. Table S3 demonstrates that we report consistently high sequence alignment performance as compared to existing approaches, even when we perform at a fully online manner only using contextual information from past frames.

S.5. Qualitative Results

We provide additional sequence alignment results of our approach in Fig. S4 and Fig. S5. Our method is able to align sequences across time by considering temporal context.

S.6. VPoser Latent Space

In Fig. S2, we observe from the t-SNE visualization of the pose embedding of VPoser [5] that the latent space is smooth and well-behaved. Different sequences with the same action are embedded in closeby locations in the embedding space and temporally close frames are mapped to nearby points in the embedding space. Therefore we conjecture that augmentations applied on the VPoser latent space would correspond to reasonable spatiotemporal augmentations in the original pose space and would enrich our dataset with diverse and realistic pose sequences.

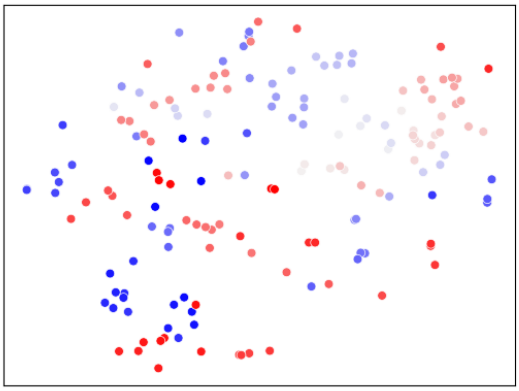

Figure S2. **t-SNE distribution.** We visualize the t-SNE distribution using *baseball_pitch* action on the PennAction dataset. Each point represents an encoded pose using VPoser [5]. We use the same color for the poses in the same sequence. In addition, we show the beginning and end frames with different shades of the same color (*e.g.* first frames are encoded with lighter colors, while later frames are encoded with darker colors.)

S.7. SVM vs Nearest Neighbor

In our main paper, we provide results for phase classification on features learned through self-supervised learning using SVM classifier. We further compute phase classification results using nearest neighbor which does not require any training data. We obtain an accuracy of 89.52% on the PennAction dataset, which is still beyond the state-of-the art, even without using any training data. Note that the previous state-of-the-art method (LAV) [3] reports 83.56%, 83.95% and 84.25% phase classification accuracy, when using an SVM classifier trained on a fraction of 10%, 50% and 100% of the ground truth labels.

S.8. Encoding Contextual Information

The global receptive field, positional encoding and self- and cross-attention layers of Transformer enable the transformed feature representations to be context- and position-dependent, as also posited by prior work [1, 7]. Therefore our architecture enables our method to be aware of the spatial and temporal context of the human actions.

S.9. Societal Impact

While our method for sequence alignment provides many beneficial use cases for AR-based task guidance, it could also be misused for surveillance and monitoring people's actions. This could raise privacy concerns and therefore use of this technology should be guided by responsible AI principles.

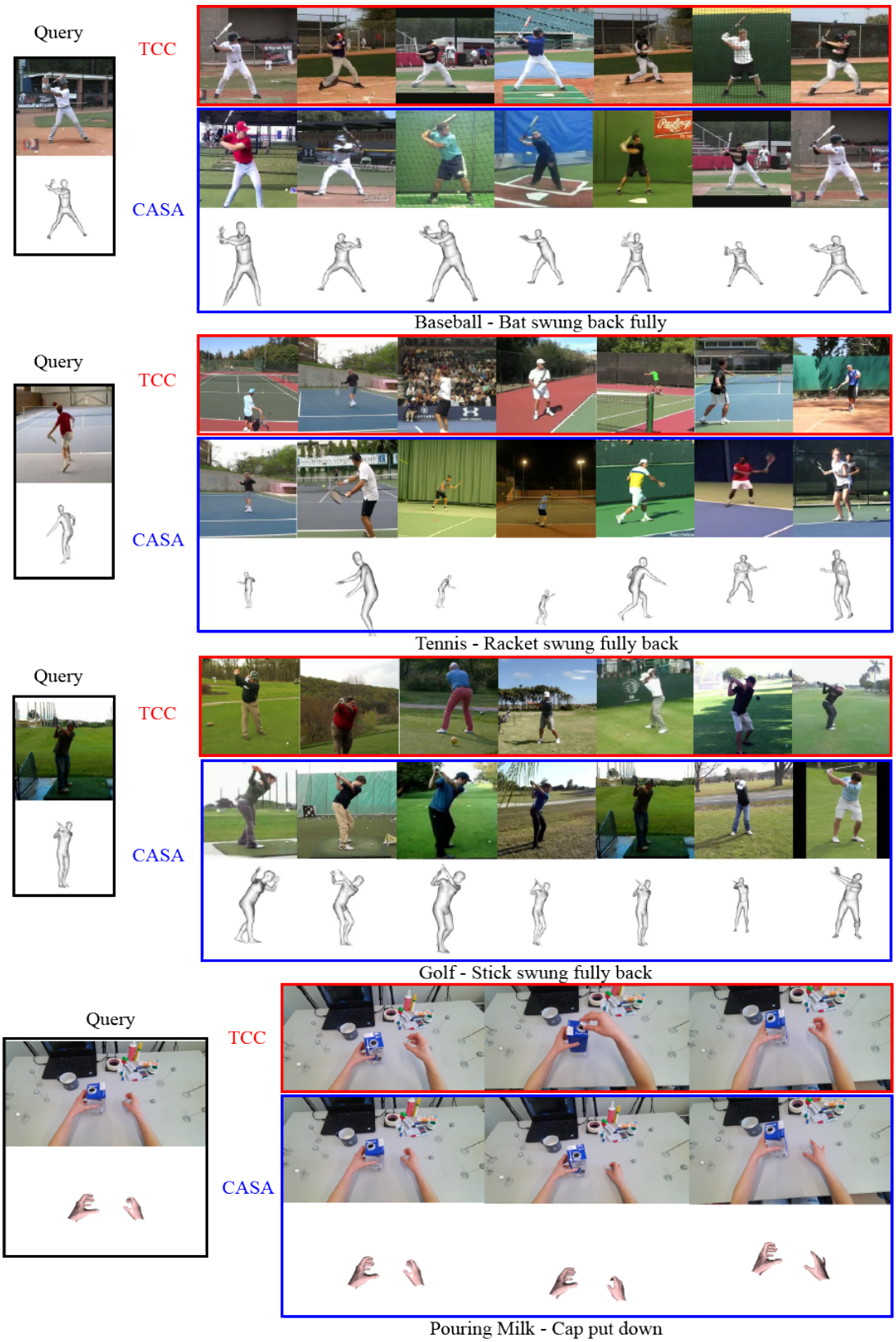

Figure S3. **Retrieval results.** We visualize our fine-grained retrieval results for the Penn Action ($k = 7$) and H2O (k=3) datasets in comparison to TCC [2] and demonstrate that our method is able to successfully retrieve visually similar frames. For example, in the "tennis" sequences, our method is able to find "racket swung fully back" action correctly in all the examples whereas TCC fails to retrieve it in some of them.

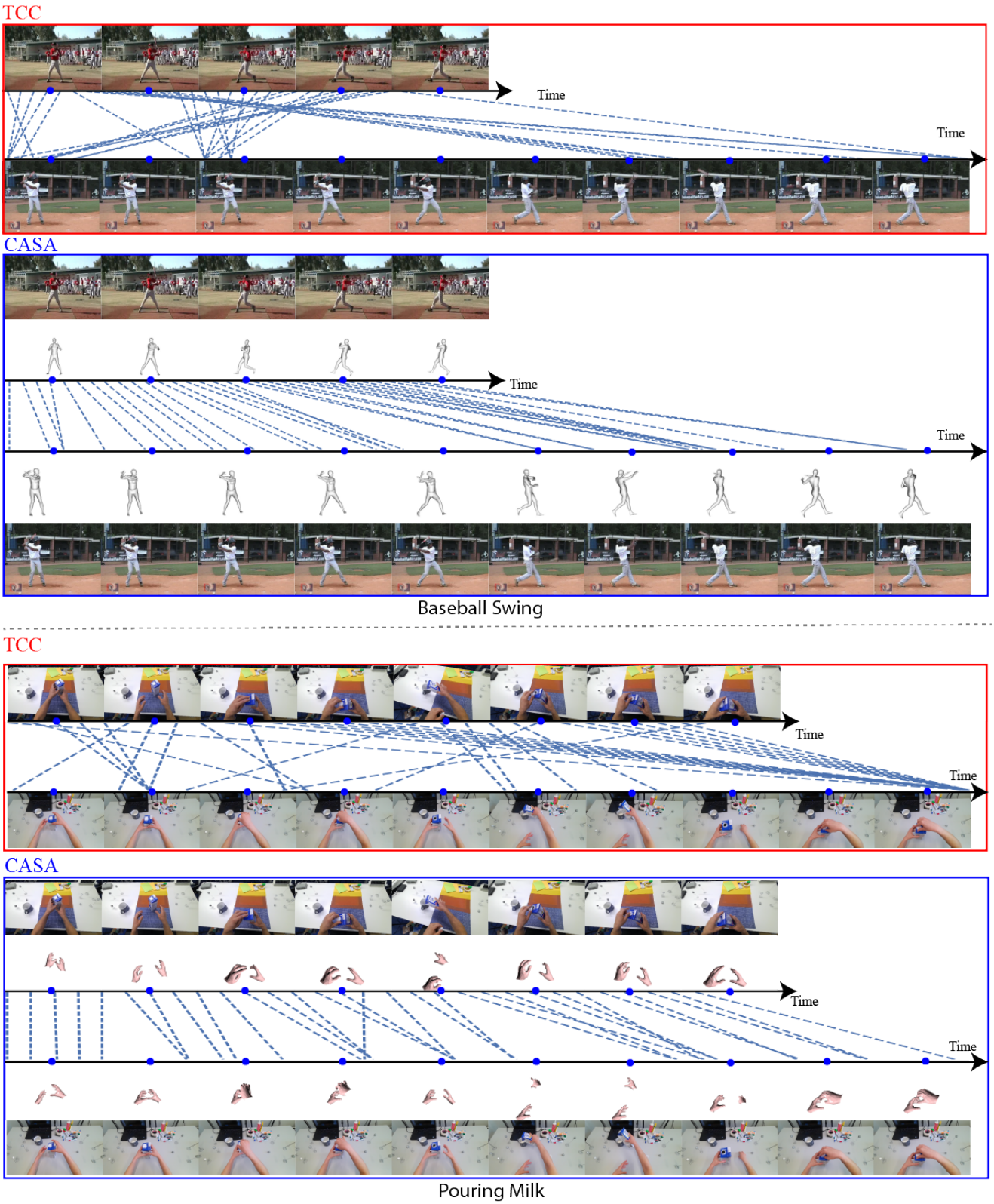

Figure S4. **Sequence alignment results**. We draw matching lines for every 20 frames in the *pouring_milk* sequence and for every frame in *baseball_swing* sequence.

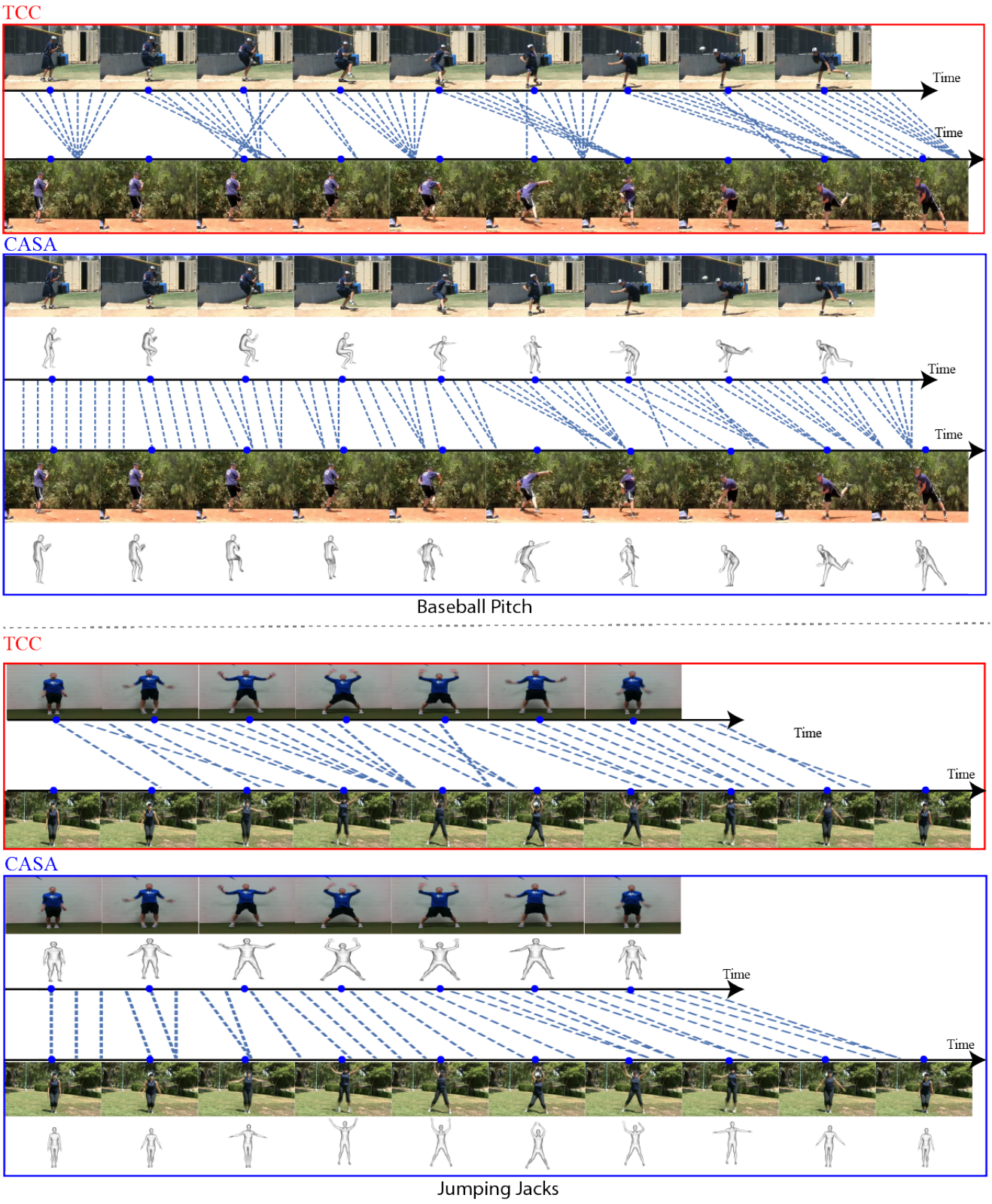

Figure S5. **Sequence alignment results**. We draw matching lines for every frame in the *baseball_pitching* and *jumping_jacks* sequences.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[2] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 1, 2, 3

[3] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021. 1, 2

[4] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2

[5] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[6] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 2

[7] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 2

[8] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2