Supplemental Material

We show in this supplemental material additional qualitative (Sec. A) and quantitative (Sec. B) results, detail our baseline evaluation protocol (Sec. C), elaborate on the 3D quality metric (Sec. D), demonstrate the ability of our method to generalize to multi-actor scenarios (Sec. E), verify our method's robustness to 2D detection results (Sec. F), show the architecture used in our approach (Sec. G), and provide additional details regarding the data (Sec. H).

A. Additional Qualitative Results

Fig. 1 shows additional qualitative results of our method, on both MPII Cooking 2 [14] (left column) and IKEA-ASM [1] (right column), as compared to pose baselines DLow [20], GSPS [12], and STARS [19].

B. Additional Quantitative Results

B.1. Characteristic Poses

Analogous to Tab. 2 in the main paper, Tab. 5 shows an ablation on pose timings and compares our approach of using characteristic poses to poses taken at regular time intervals ("uncoupled") as well as in the middle or at a random time of an action, on IKEA-ASM [1] data. To further illustrate this point, Tab. 1 shows additional ablations: Poses predicted at random points in the sequence, but at most 1s from the closest characteristic pose ("centered on the characteristic pose") and predicting characteristic poses but evaluating interpolated regularly spaced poses. Both demonstrate that the usage of characteristic poses improves performance compared to other approaches while still being outperformed by directly predicting characteristic poses.

	2D	3D	Action Accuracy	
Poses	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	75	0.29	28%	48%
Middle	58	0.45	26%	43%
Random	67	0.37	22%	42%
Centered on Char. Poses	69	0.33	28%	50%
Interp. from Char. Poses	62	0.13	29%	51%
Characteristic	**50**	**0.55**	**29%**	**51%**

Table 1. Ablation on pose forecasting on MPII Cooking II [14]. We consider pose prediction following state-of-the-art pose forecasting as decoupled from actions (uncoupled), as well as poses coupled to actions in various fashions: middle (the middle pose of an action range), random (a random pose of the action), random but at most 1s from the closest characteristic pose (centered), regularly spaced poses interpolated from characteristic pose prediction, and our characteristic pose prediction.

B.2. Lifting 2D Predictions to 3D

In Tab. 1 in the main paper, we compare to first lifting input poses into 3D, then performing 3D motion prediction. Tab. 2 evaluates the other way around: Predicting 2D poses and action labels jointly with [21], then lifting the predicted 2D poses into 3D with RepNet [17] for evaluation. Our method outperforms both approaches.

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
[21] + [17]	63	0.50	27%	43%	53	0.21	22%	46%
Ours	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	**29%**	**50%**

Table 2. Our approach of jointly forecasting 3D poses and actions achieves better performance compared to 2D pose + action forecasting [21] and then lifting forecasted 2D poses into 3D using [17].

B.3. Input Noise Ablation

Tab. 3 shows the effect using a noise vector as additional input to our method. It encourages more diversity in predictions, which benefits pose and action forecasting.

B.4. Input Objects Ablation

Inputting initially observed objects slightly improves results (Tab. 3), due to added context for broad actions like "add," e.g."add ingredient" vs. "add water to pot.".

	MPII Cooking II				IKEA ASM			
	2d	3d	Action Accuracy		2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
No Objects	61	0.52	28%	51%	42	0.30	29%	50%
No Noise	55	0.49	29%	50%	48	0.29	**30%**	**51%**
Ours	**50**	**0.55**	**29%**	**51%**	**40**	**0.31**	29%	50%

Table 3. Ablations studies with no object input and no noise input.

B.5. Statistical Action Baselines

We additionally evaluate "Zero Velocity" and "Train Average" for action labels, analogous to forecasted poses, i.e. repeating the last action label and repeating the most frequent train action label, in Tab. 4. These baselines perform particularly poorly since actions are rarely repeated or fixed for entire sequences.

	MPII Cooking II		IKEA ASM	
Approach	top-1 ↑	top-3 ↑	top-1 ↑	top-3 ↑
Repeat Last Input	9%	43%	8%	35%
Most Common in Train	6%	10%	7%	26%
Ours	**29%**	**51%**	**29%**	**50%**

Table 4. Statistical action baselines: (1) Repeat the last input action label (2) Using the most common action label of the train set.

C. Baseline Evaluation Details

C.1. State-of-the-Art Pose Forecasting

We evaluate the performance of our baselines using the same input data that is available to our method. Pose forecasting baselines DLow [20], GSPS [12], and STARS [19] are

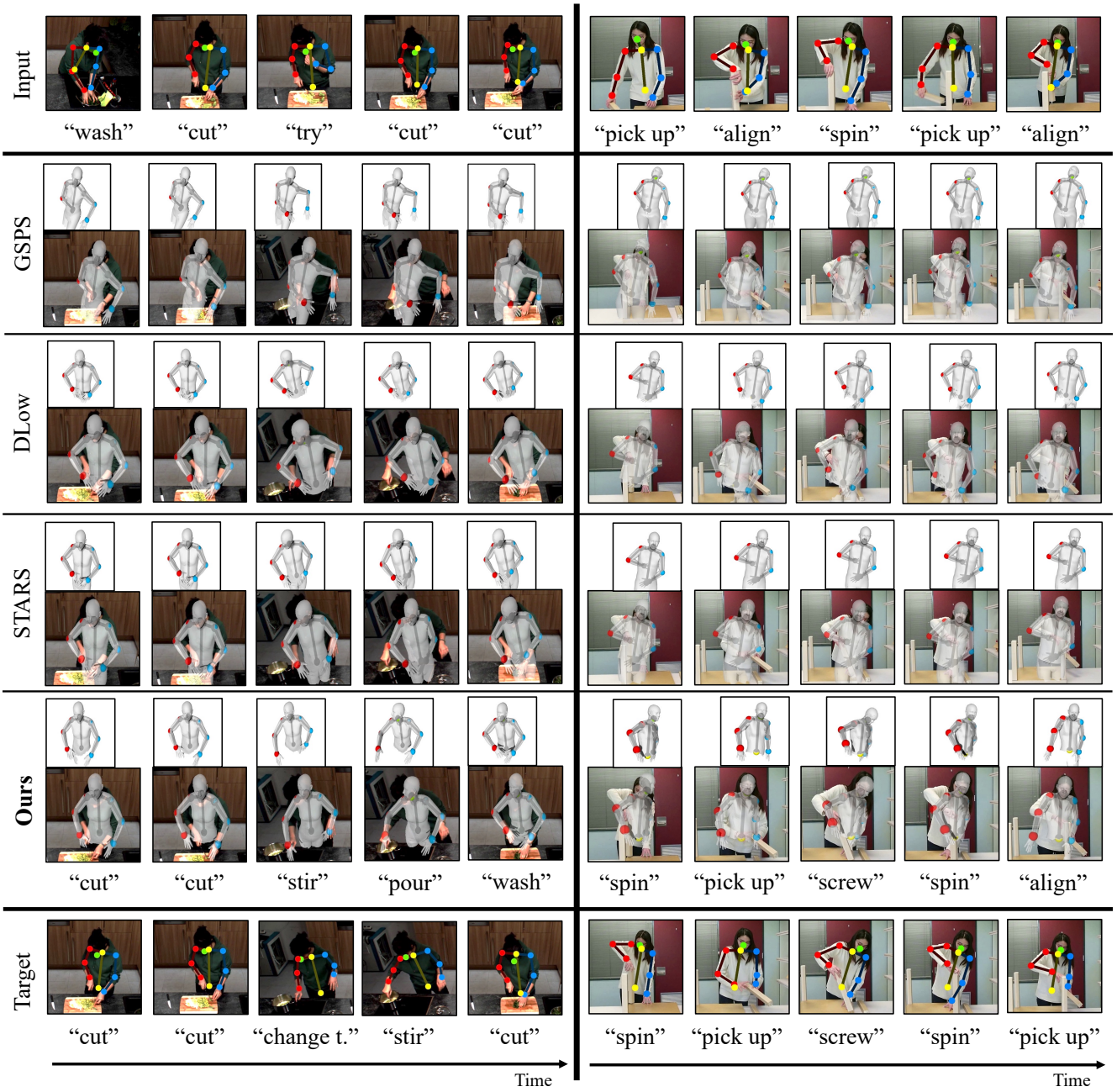


Figure 1. Additional qualitative comparison between DLow [20], GSPS [12], STARS [19], and our method on two sequences (left on MPII Cooking 2 [14], right on IKEA-ASM [1]). For each method, we show a the 3D predicted pose projected into the 2D target view, without background for a pose only version (small) as well as with background for context (full size).

trained and evaluated on sequences of our manually annotated characteristic poses. Since there is no ground-truth 3D pose data available, we first use RepNet [17], a state-of-the-art 3D pose estimation method, to retrieve 3D skeletons from our 2D characteristic poses. We train this method from scratch using the same database of valid 3D poses that is available to our method, allowing for a fair comparison.

C.2. State-of-the-Art Action Label Forecasting

We train action baselines AVT [5] and FUTR [6] using sequences of our characteristic pose frames together with the corresponding action labels as input. For AVT, we use their default parameters used by the original authors for their ablation on third-person dataset 50Salads/Breakfast, inputting our RGB frames instead. For a fair comparison, we also sup-

	2D	3D	Action Accuracy	
Poses	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
Uncoupled	64	0.30	28%	48%
Middle	47	0.35	28%	47%
Random	49	0.24	28%	49%
Characteristic	**41**	**0.35**	**29%**	**50%**

Table 5. Ablation on pose forecasting, on the IKEA-ASM [1] dataset. We consider predicting poses following state-of-the-art pose forecasting in a decoupled fashion from actions (uncoupled), as well as poses coupled to actions in various fashions: middle (the middle pose of an action range), random (a random pose of the action), and our characteristic pose prediction, which benefits action prediction the most.

ply the action and object history for each step by encoding both label sequences with a small encoder (a single linear layer) each and fuse these features with the image features generated by the AVT encoder. For FUTR, we first generate I3D features [3] from our RGB frames and concatenate them with action and object history after encoding these in the same way as for AVT.

We then train two variants of both methods: One with the raw RGB frames, action history, and object history as input ("AVT RGB" and "FUTR RGB" in the main results figure), and one with additional 2D skeleton input (skeletons rendered on top of the RGB frames) from the skeletons that we extract with OpenPose [2] ("AVT RGB+Skeleton" and "FUTR RGB+Skeleton").

C.3. Supervised 3D Pose Lifting

For better comparability, we used weakly supervised approach [17] for pose lifting. This is important, since there is no ground-truth coupling between 2D and corresponding 3D action poses in our setting. Nevertheless, we compare to baselines [12, 19, 20] in Tab. 6 with poses lifted using fully supervised pre-trained SPIN [10]; our approach outperforms even these improved baselines in terms of 2D MPJPE.

	MPII Cooking II		IKEA ASM	
	2d	3d	2d	3d
Approach	MPJPE [px] ↓	Quality ↑	MPJPE [px] ↓	Quality ↑
SPIN [10] + DLow [20]	81	**0.89**	43	**0.43**
SPIN [10] + GSPS [12]	74	0.66	45	0.29
SPIN [10] + STARS [19]	66	0.80	41	0.40
Ours	**50**	0.55	**40**	0.31

Table 6. Comparison to pose baselines using fully-supervised pre-trained 3D pose estimation method SPIN [10]. In our main experiments, we instead compare to weakly supervised baseline RepNet [17] for a fair comparison.

D. 3D Quality Metric Details

For our pose quality metric, we use a 3-layer MLP binary classifier of 3D poses. Training poses are randomly sampled from ground-truth (real) and predicted (fake) collected during the training process of our method and all baselines, producing a total of 100k real and fake poses each. Fake poses exhibit a range of small to large unrealistic deformations, depending on when they were sampled, ranging from random joint placements to widely inconsistent bone lengths to unnatural joint angles to only minor inconsistencies in the bone lengths. The classifier is trained once and then used to evaluate all methods, to ensure a fair comparison.

As an additional intuitive metric we show the mean absolute bone length difference of right and left body in 3D in Tab. 7. We observe that this metric correlates with our classifier-based quality.

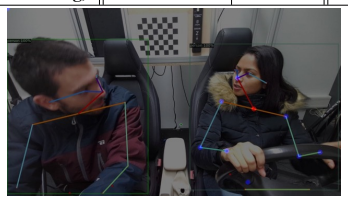
	MPII Cooking II		IKEA ASM	
Approach	Symm. [mm] ↓	Quality ↑	Symm. [mm] ↓	Quality ↑
RepNet [17] + DLow [20]	**13**	**0.72**	45	0.31
RepNet [17] + GSPS [12]	18	0.66	56	0.15
RepNet [17] + STARS [19]	16	0.62	46	0.27
No 3D Adversarial Loss	75	0.10	66	0.05
2D Projection Loss Only	57	0.21	61	0.09
No Action Loss	22	0.53	39	0.29
Ours	22	0.55	**39**	**0.31**

Table 7. Additional quality metric and its correlation to our classifier-based metric: Absolute bone length difference between right and left body, compared to pose baselines and ablations.

E. Multi-Actor Interaction Scenario

In addition to our experiments with single human actors in the main paper, we show here that our approach is able to generalize to multi-actor scenarios, with minor modifications. We show this in Tab. 8 with additional dataset TICaM [9] where driver and passenger are interacting in an in-car driving scenario (actions include "talking", various handoffs). Our modifications are: **(1)** Additional encoder and decoder for the second person **(2)** Interaction pooling introduced in Social GAN [7]. Our modified method outperforms simple combinations of previous works, with and without interaction modelling, demonstrating the wide applicability of our method.

	2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
FUTR RGB + Skeleton	-	-	38%	64%
RepNet + STARS	89	0.34	-	-
Ours (No Interactions)	68	0.40	40%	67%
Ours (Interaction Modeling)	**58**	**0.41**	**48%**	**73%**



Setting

Table 8. Our approach can also be applied to multi-actor scenarios: We demonstrate improved performance on suitable dataset TICaM [9], with and without explicit interaction modeling.

F. 2D Input Pose Quality

In Fig. 9, we replace OpenPose with AlphaPose [4] and Detectron2 [18], both only slightly changing the final results, indicating that our method does not depend on a specific 2D pose detector. We also experiment with added random noise to OpenPose: our method remains relatively robust. The coupled changes in pose and action accuracy further demonstrate the effectiveness of our joint feature learning.

MPII Cooking II	2d	3d	Action Accuracy	
Approach	MPJPE [px] ↓	Quality ↑	top-1 ↑	top-3 ↑
OpenPose + max. 20px noise	59	0.45	26%	47%
OpenPose + max. 10px noise	57	0.47	26%	46%
Ours (using Detectron2)	47	**0.54**	28%	55%
Ours (using AlphaPose)	**46**	0.57	28%	**56%**
Ours (using OpenPose)	50	0.55	**29%**	51%

Table 9. Robustness of our method to different 2D pose detectors Detectron2 [18] and AlphaPose [4] as well as randomly added 2D noise. This only slightly affects our pose and action accuracy, further demonstrating the effectiveness of our joint feature learning.

G. Architecture Details

Generator Network Fig. 2 shows our generator architecture in detail with input and output dimensions for linear layers, and the slope for leaky ReLU layers.

Critic Network Our adversarial critic network processes generator outputs with 4 linear layers and 3 kinematic chain layers which are designed to encourage correct bone lengths (as shown in [17]), in parallel. 2 linear layers then combine both outputs and produce the final critic score.

H. Data Details

H.1. Camera Parameters

While intrinsic camera parameters are often available in captured image data, the camera parameters for captured video were not available from the MPII Cooking 2 [14] dataset to use for pose projection. We thus optimized for intrinsic camera parameters from the video sequence data in correspondence with the 3D scene reconstruction of the empty kitchen environment, as given by [15]. For IKEA-ASM [1], we use the provided intrinsic camera parameters directly. Note that camera parameters are only required to be fixed within a sequence (i.e. no moving camera) but can change between sequences.

H.2. 3D Pose Database Alignment

We use popular 3D pose datasets Human3.6m [8], AMASS [11], and GRAB [16] for our database of uncorrelated valid 3D poses. All poses are pre-processed to follow the OpenGL coordinate system and aligned with respect to the neck joint.

Ours		OpenPose		Human3.6m		SMPL-X	
Idx	Name	Idx	Name	Idx	Name	Idx	Name
0	head	0	nose	15	head	15	head
1	neck	1	neck	13	thorax	12	neck
2	right shoulder	2	right shoulder	25	right shoulder	17	right shoulder
3	right elbow	3	right elbow	26	right elbow	19	right elbow
4	right hand	4	right hand	27	right wrist	42	right index 3
5	left shoulder	5	left shoulder	17	left shoulder	16	left shoulder
6	left elbow	6	left elbow	18	left elbow	18	left elbow
7	left hand	7	left wrist	19	left wrist	27	left index 3
8	hip	8	mid-hip	0	hip	0	pelvis

Table 10. Human skeleton joint layout used in our experiments, for both 2D and 3D skeletons.

H.3. Pose Joint Layout

We use the 9 upper-body joints of the native OpenPose [2] joint layout for skeletons in 2D, and adapt skeletons in our 3D database to use the same format. Tab. 10 shows the correspondence between our joint layout, OpenPose [2], Human3.6m [8], and SMPL-X [13]. 3D datasets AMASS [11] and GRAB [16] provide human bodies in SMPL-X format; we first extract their skeleton joints using their publicly available code and then convert it into our layout using the correspondences in Tab. 10.

H.4. MPII Cooking 2 Details

We use action labels as annotated in the 2D cooking action dataset MPII Cooking 2 [14]. These annotations provide action labels (87 classes) for frame ranges in each sequence as well as the involved objects (187 classes). We first cluster similar actions together, yielding a total of 37 action clusters, which we then use as action classes in our experiments.

In addition, since our goal is to forecast upper-body actions with objects in the foreground, we remove instances of poses and corresponding actions that occur in the background - e.g., when taking out objects from the cupboard, or from the fridge.

In total, there are 272 cooking action sequences; we create a random train/val/test split along sequences with a ratio of 70% / 15% / 15%, yielding 190, 40, 40 sequences for each set.

H.5. IKEA-ASM Details

We use action labels as annotated in the IKEA furniture assembly dataset IKEA-ASM [1]. These annotations provide action labels (31 classes) for frame ranges in each sequence; we use them without explicit object information since each action already encodes its associated object.

In total, there are 370 furniture assembly action sequences; we create a random train/val/test split along sequences with a ratio of 70% / 15% / 15%, yielding 227, 48, 48 sequences for each set.

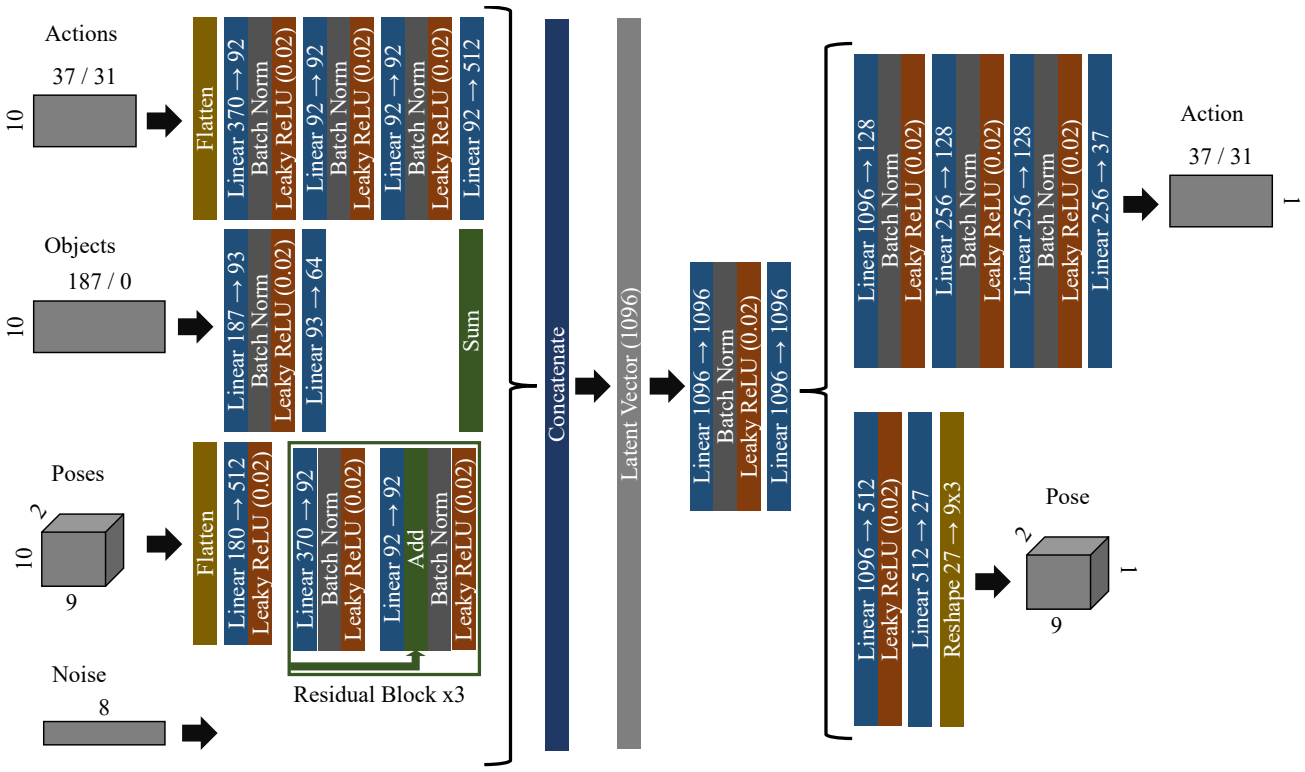


Figure 2. Network architecture specification.

References

[1] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 1, 2, 3, 4

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 4

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017. 3

[4] Haoshu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45 (6):7157–7173, 2023. 4

[5] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13485–13495. IEEE, 2021. 2

[6] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3042–3051. IEEE, 2022. 2

[7] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: socially acceptable trajectories with generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2255–2264. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 4

[9] Jigyasa Singh Katrolia, Ahmed El-Sherif, Hartmut Feld, Bruno Mirbach, Jason R. Rambach, and Didier Stricker. Ticam: A time-of-flight in-car cabin monitoring dataset. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 277. BMVA Press, 2021. 3

[10] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2252–2261. IEEE, 2019. 3

[11] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Ger-

ard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 4

[12] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021. 1, 2, 3

[13] 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)*, pages 10975–10985, 2019. 4

[14] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, pages 1–28, 2015. 1, 2, 4

[15] Wandi Susanto, Marcus Rohrbach, and Bernt Schiele. 3d object detection with multiple kinects. In *European Conference on Computer Vision*, pages 93–102. Springer, 2012. 4

[16] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision – ECCV 2020*, pages 581–600, Cham, 2020. Springer International Publishing. 4

[17] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019. 1, 2, 3, 4

[18] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[19] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXII*, pages 251–269. Springer, 2022. 1, 2, 3

[20] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 1, 2, 3

[21] Yanjun Zhu, David Doermann, Yanxia Zhang, Qiong Liu, and Andreas Girgensohn. What and how? jointly forecasting human action and pose. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 771–778. IEEE, 2021. 1