Pre-Training for 3D Hand Pose Estimation with Contrastive Learning on Large-Scale Hand Images in the Wild

Nie Lin*, Takehiko Ohkawa*, Mingfang Zhang,
Yifei Huang, Ryosuke Furuta, Yoichi Sato
Institute of Industrial Science
The University of Tokyo
{nielin, ohkawa-t, mfzhang, hyf, furuta, ysato}@iis.u-tokyo.ac.jp

Abstract

We present a contrastive learning framework based on in-the-wild hand images tailored for pre-training 3D hand pose estimators, dubbed HandCLR. Pre-training on large-scale images achieves promising results in various tasks, but prior 3D hand pose pre-training methods have not fully utilized the potential of diverse hand images accessible from in-the-wild videos. To facilitate scalable pre-training, we first prepare an extensive pool of hand images from in-the-wild videos and design our method with contrastive learning. Specifically, we collected over 2.0M hand images from recent human-centric videos, such as 100DOH and Ego4D. To extract discriminative information from these images, we focus on the similarity of hands; pairs of similar hand poses originating from different samples, and propose a novel contrastive learning method that embeds similar hand pairs closer in the latent space. Our experiments demonstrate that our method outperforms conventional contrastive learning approaches that produce positive pairs sorely from a single image with data augmentation. We achieve significant improvements over the state-of-the-art method in various datasets, with gains of 15% on FreiHand, 10% on DexYCB, and 4% on AssemblyHands.

1. Introduction

Hands are a trigger for us to interact with the world, as seen in various human-centric videos. Precise recognition of hand states, such as 3D keypoints, is crucial for video understanding [34, 39], AR/VR interfaces [18, 40], and robot learning [6, 31]. To this end, 3D hand pose estimation has been studied through constructing labeled datasets [6, 26, 27, 45] and advancing pose estimators [3, 12, 13, 22, 30]. However, utilizing large-scale, unannotated hand videos for pre-training remains underexplored, while vast collections of such videos, like 3,670 hours of videos from Ego4D [16] and 131 days from 100DOH [35], are available.

Some works utilize unlabeled hand images for 3D hand pose pre-training using contrastive learning like SimCLR [9], which maximizes agreement between positive pairs while repelling negatives. Spurr *et al*. [38] introduce pose equivariant contrastive learning (PeCLR) by aligning geometry in latent space after affine transformations for input images. However, both SimCLR and PeCLR create positive pairs from a single sample by applying augmentation, limiting the gains from positive pairs as their hand appearance and backgrounds are identical. Ziani *et al*. [44] extend the contrastive learning framework to video sequences by treating temporally adjacent frames as positive pairs. However, in-the-wild videos can challenge tracking hands across frames, especially in egocentric views where hands may be unobservable due to camera motion. In addition, adjacent frames still pose limited appearance variation of hands and backgrounds.

In this work, we introduce a novel contrastive learning framework for 3D hand pose pre-training to leverage diverse hand images in the wild, with the largest 3D hand pose pre-training set to date. We collected 2.0M hand images from in-the-wild videos, specifically from Ego4D [16] and 100DOH [35], using an off-the-shelf hand detector [24]. Our pre-training set significantly exceeds the scale of prior works by two orders of magnitude, such as the 32-47K images in [38] and 86K images from 100DOH in [44].

Our method focuses on learning discriminative information by leveraging the similarity of hands from different domains. Unlike SimCLR and PeCLR, we observe that it is further informative to learn from positive pairs with similar foreground hands but from different images. As shown in Fig. 1, our positive pairs based on different images offer additional information gains from different types of object interactions, backgrounds, and hand appearances. Specifically, we use an off-the-shelf 2D hand pose estimator [24] to identify similar hands from the pre-training set.

*Equal contribution

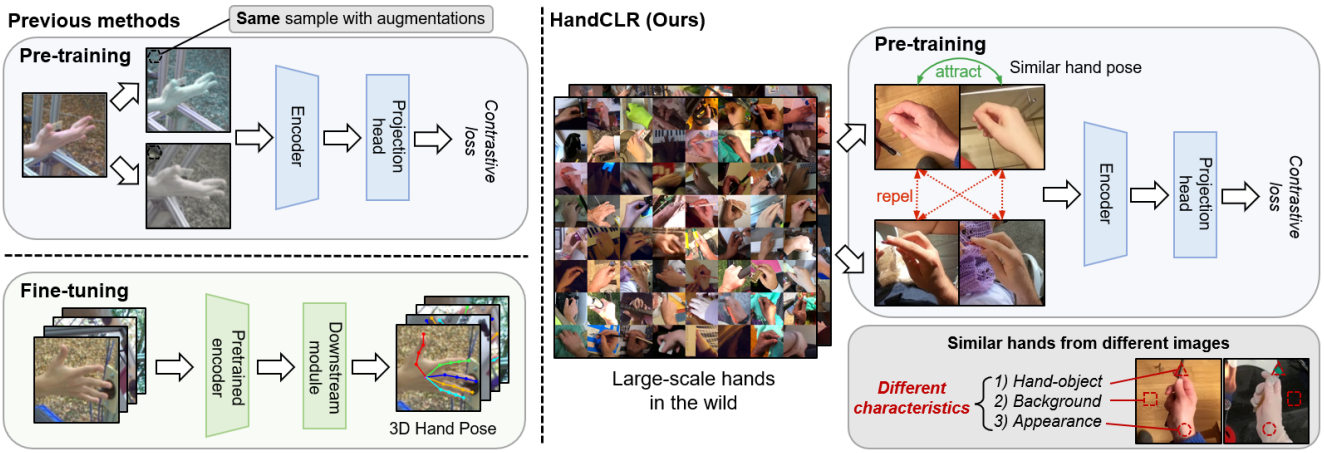

Figure 1. **The pipeline of pre-training and fine-tuning in 3D hand pose estimation. (Left)** Previous pre-training methods (*e.g.*, PeCLR [38]) learn from positive pairs originating from the same with different augmentations and fine-tune the network on a dataset. **(Right)** Our method is designed to learn from positive pairs with similar foreground hands, sampled from a pool of hand images in the wild.

Our contributions are threefold: 1) We construct a large-scale in-the-wild hand dataset for 3D hand pose pre-training. 2) We propose a pre-training framework using similar hand pairs via contrastive learning. 3) Our model achieves state-of-the-art performance across multiple datasets.

2. Related Work

3D hand pose estimation: The task of 3D hand pose estimation aims to regress 3D keypoints of hand joints. Annotating 3D hand poses is challenging, which allows us to have limited labeled datasets [26], mostly constructed in controlled laboratory settings [6, 25, 27, 45]. Given this challenge, two approaches have been proposed to facilitate learning from limited annotations: pseudo-labeling [22, 23, 28, 41] and self-supervised pre-training [38, 44]. Pseudo-labeling methods learn from pseudo-ground-truth assigned on unlabeled images [22, 23, 28, 41]. Alternatively, pre-training methods first pre-train an encoder with contrastive learning on unlabeled images and then fine-tune on labeled images [38, 44]. While prior works use relatively small pre-training sets (*e.g.*, 32-47K images in [38] and 86K images in [44]), our work emphasizes leveraging in-the-wild images on a large scale. We collected hand images from large human-centric datasets such as Ego4D [16] and 100DOH [35], expanding our pre-training set to 2.0M images.

Contrastive learning: The framework of contrastive learning has emerged as a powerful self-supervised learning, bringing positive samples closer while pushing negative samples apart [11, 20, 29, 33, 36, 37]. Standard methods generate positive samples from individual images with data augmentation (*i.e.*, self-positives) [4, 5, 10, 21, 32], but restrict the learning of explicit relationships between samples. To address this, Zhang *et al*. propose a relaxed extension of self-positives, *non-self-positives* [42], which share similar characteristics (*e.g.*, same scene [1, 2, 14, 19] or instance [7, 8, 15, 17]) but originate different images. This enables the incorporation of diverse inter-sample consistency and facilitates the learning of semantics more easily. Skeleton-based action recognition methods identify non-self-positives by searching similar human skeletons [42], whereas it relies on online mining, increasing computational overhead in training. In contrast, our approach creates non-self-positives using 2D hand keypoints offline for avoiding the overhead and scaling pre-training with any large data.

3. Method

Our approach called HandCLR aims to pre-train an encoder of a 3D hand pose estimator with large human-centric videos available in the wild. We first construct the pre-training set from egocentric and exocentric hand videos (Sec. 3.1), then find similar hand images to define positive pairs (Sec. 3.2), and finally incorporate these positive pairs into a contrastive learning framework (Sec. 3.3).

3.1. Data preprocessing

Our preprocessing involves creating a set of valid hand images for pre-training, which is sampled from a dataset with N videos, $\{v_1, v_2, \ldots, v_N\}$. We use an off-the-shelf hand detector [35] to select valid frames with hands. Given an image of a video, $I_{\text{full}} \in v_i$, the model detects the existence of the hand and gives hand crops enclosing either hand identity (right/left) from I_{full}. To avoid bias regarding hand identity, we balance the number of right and left hand crops and convert all crops to the right, allowing us to address all crops equally. Then, we create a frame set for each video as $\mathcal{F}_i = \{I_{i,1}, I_{i,2}, \ldots, I_{i,T_i}\}$, where $I_{i,j} \in \mathbb{R}^{H \times W \times 3}$ represents the processed crop with height H and width W,

and T_i is the number of crops in the video v_i. The height H and width W are defined post-resize to give the uniform image size. Using this frame set $\mathcal{F}_i$, the video dataset can be re-represented as $\mathcal{V} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$. Specifically, we processed two datasets, Ego4D [16] and 100DOH [35], comprising 8K and 21K videos, respectively.

3.2. Mining similar hands

Our preliminary experiments indicate that learning from positive pairs with similar foreground hands but from different images could provide additional information gains compared to conventional contrastive learning [9, 38]. Here we construct a mining algorithm for similar hands from $\mathcal{V}$ by focusing on pose similarity between hand images. We first extract 2D pose from I, embedding in the latent space, and design a scheme for effective positive sample mining.

Pose embedding: To compute the hand pose similarity robustly, we obtain a D-dimensional embedding of 2D hand keypoints, $\mathbf{p} \in \mathbb{R}^D$, for each image I. Using an off-the-shelf 2D hand pose estimator ϕ [24], we predict 2D keypoints for 21 joints. We use a concatenated 42-dimensional vector as the output of ϕ for later use. As these 2D keypoints are prone to be noisy, we apply PCA-based dimensionality reduction to project the vector into a lower-dimensional space of size D. Given the PCA projection matrix $M \in \mathbb{R}^{42\times D}$, the pose embedding $\mathbf{p}$ is calculated as $\mathbf{p} = M^T\phi(I)$. This process mitigates noise and provides a more robust representation. We empirically choose $D = 14$ for our experiments.

Mining: This step is designed to identify a positive sample $J \in \mathbb{R}^{H\times W\times 3}$ paired with a query image I. We denote the similarity mining logic as $J = \mathrm{SiM}(I)$. When using the closest sample in the PCA space, we encounter a trivial solution $I, J \in v_i$, where both images originate from the same video v_i. Similarly to [44], the supervision by positive samples from the same video have less diversity in backgrounds, hand appearances, and object interactions. Thus we are motivated to find similar hands derived from different video domains. Specifically, we search the minimum distance within the set of all frames except for v_i, written as $\mathcal{F}_{\mathrm{excl},i} = \bigcup_{\substack{k=1\\k\neq i}}^{N}\bigcup_{j=1}^{T_k} I_{kj}$. Given an query $I_{i,j}$ where a j-th image of a i-th video, the function $\mathrm{SiM}(\cdot)$ is formulated as

$$\mathrm{SiM}(I_{i,j}) = \arg\min_{x\in\mathcal{F}_{\mathrm{excl},i}} d(M^T\phi(x), M^T\phi(I_{i,j})), \tag{1}$$

where $d(\cdot,\cdot)$ is the Euclidean distance metric.

3.3. Contrastive learning

Given the positive samples (I, J) constructed by Sec. 3.2, we describe feature extraction process and contrastive learning loss. Following [9, 38], we treat all samples other than its corresponding positive sample as negative samples. In our framework, feature extraction is performed by two learnable components: an encoding head $E(\cdot)$ and a projection head $g(\cdot)$. We define image augmentation as $\mathbf{T}$ and the entire model as $f = g \circ E$. Given the positive pair (I, J), feature extraction is performed as $\mathbf{z} = f(\mathbf{T}(I))$ and $\mathbf{z}^+ = f(\mathbf{T}(J))$. Since $\mathbf{T}$ introduces geometric transformations that may cause the misalignment between the image and the feature spaces, we correct such error with the inverse transformation $\mathbf{T}^{-1}$ as [38]. Finally, we use the *NT-Xent* loss [9] for contrastive learning, enabling the feature alignment between $\mathbf{z}$ and $\mathbf{z}^+$. For fine-tuning, we initialize our model with the pre-trained encoder $E(\cdot)$ and then fine-tune with a 3D pose regressor on labeled datasets. The 3D regressor involves 2D heatmap regression and 3D localization, inspired by DetNet [43].

4. Experiments

In this section, we begin by detailing the datasets and present our key experiments by comparing our results with state-of-the-art methods.

4.1. Datasets

Pre-training datasets: We collected a large collection of hand images from two major video datasets, Ego4D [16] and 100DOH [35], capturing egocentric and exocentric views respectively. From Ego4D, a vast egocentric video dataset with 3,670 hours of footage, we extracted 1.0M hand images from 8K videos. Similarly, from the exocentric dataset 100DOH, which includes 131 days of YouTube footage and 100K annotated hand-object interaction frames, we extracted 1.0M hand images from 20K videos. These extensive datasets provide diverse hand-object interactions across different views.

Fine-tuning datasets: We conduct fine-tuning experiments on three datasets with ground-truth 3D hand pose: FreiHand [45], DexYCB [6], and AssemblyHands [27]. FreiHand, with 130K training frames, includes both green screen and real-world backgrounds, while DexYCB offers 582K images of natural hand-object interactions in a controlled laboratory setting. AssemblyHands, the largest of the three, consists of 412K training and 62K test samples, collected from egocentric perspectives in object assembly scenarios.

4.2. Results

As shown in Tab. 1, we compare our method with state-of-the-art pre-training methods for 3D hand pose estimation using the metrics of MPJPE (↓) and PCK-AUC (↑).

Pre-training results: We observe that our method significantly outperforms SimCLR and PeCLR across various datasets under the same pre-training data setup. Specifically, on the FreiHand dataset, our approach achieves a 15.3% improvement with Ego4D-1M pre-training. Furthermore, our method demonstrates strong performance on larger datasets, with a 10.53% gain on DexYCB and a 4.90% improvement on AssemblyHands compared to PeCLR. These results con-

Method	Pre-training	FreiHand (Exo) [45]		DexYCB (Exo) [6]		AssemblyHands (Ego) [27]	
		MPJPE↓	PCK-AUC↑	MPJPE↓	PCK-AUC↑	MPJPE↓	PCK-AUC↑
SimCLR [9]	100DOH-1M	19.30	85.36	20.13	83.75	20.01	84.21
	Ego4D-1M	19.36	85.09	20.22	83.50	20.32	83.85
PeCLR [38]	100DOH-1M	19.58	84.71	18.39	18.39	19.12	85.64
	Ego4D-1M	19.07	85.62	18.99	85.40	19.20	85.57
HandCLR	100DOH-1M	16.73	88.66	17.34	87.84	18.50	86.56
	Ego4D-1M	16.15	89.48	16.99	88.34	18.26	86.95
	Ego4D-1M+100DOH-1M	**15.79**	**90.04**	**16.71**	**88.86**	**18.23**	**86.90**

Table 1. **Comparison with the state of the art.** We show 3D hand pose estimation accuracy (MPJPE↓) on the FreiHand (Exo) [45], DexYCB (Exo) [6] and AssemblyHands (Ego) [27]. Our method achieves the best results across various pre-training datasets.

Pre-training size	Method	MPJPE ↓	PCK-AUC ↑
Ego4D-50K	SimCLR	53.94	42.54
	PeCLR	47.42	49.85
	HandCLR	**35.32**	**63.35**
Ego4D-100K	SimCLR	53.49	43.12
	PeCLR	46.00	51.50
	HandCLR	**31.06**	**68.66**
Ego4D-500K	SimCLR	49.91	47.61
	PeCLR	43.18	54.15
	HandCLR	**28.27**	**72.97**
Ego4D-1M	SimCLR	46.17	50.62
	PeCLR	34.42	64.93
	HandCLR	**23.68**	**79.62**

Table 2. **Comparison with different pre-training data sizes.** We use 10% of the labeled FreiHand [45] dataset for fine-tuning.

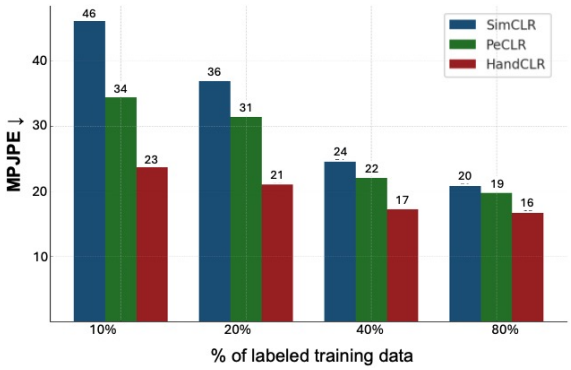

Figure 2. **Comparison with different data availability in fine-tuning.** Variations in the percentage of labeled data correspond to different subsets of the FreiHand [45] dataset, following the experimental design in [38].

firm that our model consistently achieves superior performance across various pre-training datasets.

Performance on Ego & Exo hands: We evaluate how pre-training with egocentric views (Ego4D) and exocentric views (100DOH) affects the performance in datasets with their corresponding views, namely AssemblyHands for egocentric and FreiHand and DexYCB for exocentric views. Interestingly, matching pre-training viewpoints does not consistently enhance performance, indicating that the view gaps have limited effects. Instead, factors like dataset diversity and the characteristics of pre-training methods are more crucial in determining effectiveness. We also assess pre-training performance using both perspectives, Ego4D and 100DOH. Combining the two datasets, the last row of Tab. 1, leads to the best performance in all three datasets, underscoring the potential of enriching data diversity with different camera characteristics.

Effect of different pre-training data sizes: We study results with different sizes of pre-training data, namely 50K, 100K, 500K, and 1M in Tab. 2. We specifically test the pre-trained networks on limited labeled data, *i.e.*, 10% of FreiHand. This shows that HandCLR consistently improves in all settings, with gains increasing further with more pre-training data.

Results in smaller fine-tuning sets: Fig. 2 illustrates the MPJPE performance comparison of three methods under different proportions of labeled data, namely 10%, 20%, 40%, and 80% in FreiHand. The results show that our HandCLR method performs particularly well in a limited data regime, such as 10% and 20%, compared to the baselines.

5. Conclusion

We introduce a contrastive learning framework for pre-training 3D hand pose estimators using the largest in-the-wild pre-training set. Our approach leverages similar hand pairs from diverse videos, significantly enhancing the information gained during pre-training over existing methods. Experiments show our method achieves state-of-the-art performance in 3D hand pose estimation across multiple datasets. This work demonstrates the benefits of pre-training with large-scale in-the-wild images and lays the foundation for future research on using diverse human-centric videos to improve the robustness of 3D hand pose estimation.

Acknowledgments: We thank Minjie Cai for helpful discussions on this manuscript. This work was supported by the JST ACT-X Grant Number JPMJAX2007, JST ASPIRE Grant Number JPMJAP2303, and JSPS KAKENHI Grant Number 24K02956.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016. 2

[2] G. Berton, C. Masone, and B. Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4878–4888, 2022. 2

[3] Y. Cai, L. Ge, J. Cai, and J. Yuan. Weakly-supervised 3D hand pose estimation from monocular RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 678–694, 2018. 1

[4] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 9912–9924, 2020. 2

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2

[6] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 1, 2, 3, 4

[7] H. Chen, B. Lagadec, and F. Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14960–14969, 2021. 2

[8] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2004–2013, 2021. 2

[9] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 1, 3, 4

[10] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 2

[11] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005. 2

[12] Z. Fan, T. Ohkawa, L. Yang, N. Lin, Z. Zhou, S. Zhou, J. Liang, Z. Gao, X. Zhang, X. Zhang, F. Li, L. Zheng, F. Lu, K. A. Zeid, B. Leibe, J. On, S. Baek, A. Prakash, S. Gupta, K. He, Y. Sato, O. Hilliges, H. J. Chang, and A. Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[13] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3D hand shape and pose estimation from a single RGB image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10833–10842, 2019. 1

[14] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–386, 2020. 2

[15] Y. Ge, F. Zhu, D. Chen, and R. Zhao. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 11309–11321, 2020. 2

[16] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M.g Xu, E. Zhongcong Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. Soo Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, Lo Torresani, M.i Yan, and J. Malik. Ego4D: Around the world in 3, 000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2022. 1, 2, 3

[17] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, and J. Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19243–19253, 2023. 2

[18] S. Han, P.-C. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan, R. Cabezas, L. Tran, M. Akbay, T.-H. Yu, C. Keskin, and R. Wang. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *In Proceedings of the ACM SIGGRAPH Asia Conference*, pages 50:1–50:9, 2022. 1

[19] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14141–14152, 2021. 2

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2

[21] F. Altché C. Tallec P. Richemond E. Buchatskaya C. Doersch B. Avila Pires Z. Guo M. Gheshlaghi Azar B. Piot

K. Kavukcuoglu R. Munos M. Valko J. Grill, F. Strub. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 21271–21284, 2020. 2

[22] R .Liu, T. Ohkawa, M. Zhang, and Y. Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 677–686, 2024. 1, 2

[23] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2021. 2

[24] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, and F. Zhang et al. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 1, 3

[25] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 548–564, 2020. 2

[26] T. Ohkawa, R. Furuta, and Y. Sato. Efficient annotation and learning for 3D hand pose estimation: A survey. *International Journal on Computer Vision (IJCV)*, 131:3193—3206, 2023. 1, 2

[27] T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 1, 2, 3, 4

[28] T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–87, 2022. 2

[29] A. Oord, Y. Li, and O .Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2

[30] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1495, 2022. 1

[31] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–587, 2022. 1

[32] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 2

[34] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21096–21106, 2022. 1

[35] D. Shan, J. Geng, M. Shu, and D. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9866–9875, 2020. 1, 2, 3

[36] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1857–1865, 2016. 2

[37] H. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016. 2

[38] A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges. Self-supervised 3D hand pose estimation from monocular RGB via contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11210–11219, 2021. 1, 2, 3, 4

[39] Y. Wen, H. Pan, T. Ohkawa, L. Yang, J. Pan, Y. Sato, T. Komura, and W. Wang. Generative hierarchical temporal transformer for hand action recognition and motion prediction. *CoRR*, abs/2311.17366, 2023. 1

[40] M.-Y. Wu, P.-W. Ting, Y.-H. Tang, E. T. Chou, and L.-C. Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation*, 70:102802, 04 2020. 1

[41] L. Yang, S. Chen, and A. Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11364–11373, 2021. 2

[42] H. Zhang, Y. Hou, W. Zhang, and W. Li. Contrastive positive mining for unsupervised 3d action representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–51, 2022. 2

[43] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5346–5355, 2020. 3

[44] A. Ziani, Z. Fan, M. Kocabas, S. J. Christen, and O. Hilliges. Tempclr: Reconstructing hands via time-coherent contrastive learning. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 627–636, 2022. 1, 2, 3

[45] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. J. Argus, and T. Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 1, 2, 3, 4