Supplementary Material:
Towards Egocentric 3D Hand Pose Estimation in Unseen Domains

Table 7. **Results of 3D hand pose estimation** state-of-the-art methods when trained and tested within the same domain - *H2O Dataset*. All numbers provided in *mm* in camera space.

Method	Year	MPJPE ↓
LPC [5]	2020	40.72
H+O [14]	2019	40.14
H2O [6]	2021	39.33
HTT [15]	2023	35.33
H2OTR [3]	2023	25.10
THOR-Net [1]	2023	36.65
SHARP [8]	2024	28.66
Ours	**2025**	**22.77**

A. Additional Results and Experiments

A.1. Comparison of V-HPOT in H2O Dataset as source-target

To evaluate our model's effectiveness relative to current state-of-the-art approaches, we conducted same-domain experiments using the *H2O Dataset* [6] for both training and testing. The results demonstrate that our model outperforms existing methods. The second-best performing approach, H2OTR [3], employs a substantially more complex architecture incorporating hand-object contact maps. Yet, our more streamlined design still yields better results (See Table 7). Our solution's strong performance is due to the 3D augmentation enabled by virtual camera space.

A.2. Dataset selection and cross-dataset evaluation

We train our model on *HOT3D* [2] and evaluate it on the *H2O Dataset* [6] and *AssemblyHands* [9]. We select *HOT3D* as the training dataset for two key reasons: (1) it provides the most recent high-quality motion capture data with diverse scene variations and (2) it enables direct comparison with existing methods using *H2O* and *AssemblyHands* for cross-domain evaluation [3, 10, 11, 13]. Table 8 presents a comprehensive cross-dataset ablation study demonstrating that V-HPOT consistently improves average performance across all training–testing combinations. Training on *HOT3D* achieves the highest average improvement on unseen domains (56.5%), significantly outperforming *AssemblyHands* (11.8%) or *H2O* (6.4%). This superior performance is a testament to the high quality of the motion capture data and the diverse environmental conditions in *HOT3D*, which provide better domain priors than the smaller *H2O* and *AssemblyHands*, with their strongly distorted monochromatic images. Our approach focuses specifically on cross-domain generalisation: training on one domain and testing on others. Unlike the state-of-the-art methods presented in Table 6, which rely on multiple training domains through purely data-driven approaches, V-HPOT proposes a methodological solution that requires no additional data or labels. This design choice aligns with our goal of improving performance in unseen domains while minimising the quantity of data and domains.

Table 8. **Cross-dataset 3D pose estimation performance (MPJPE in mm).** Green/red values show percentage improvement/degradation from baseline. The right column displays average improvement across test datasets, demonstrating that V-HPOT consistently improves performance in all cross-dataset scenarios.

	Train	Test		Avg. Δ
	HOT3D	H2O	AsHa	
base		179.6	297.7	–
V-HPOT		53.3 (-70.3%)	174.5 (-41.4%)	-55.9%
	AsHa	H2O	HOT3D	
base		252.1	349.6	–
V-HPOT		204.2 (-19.0%)	333.3 (-4.6%)	-11.8%
	H2O	AsHa	HOT3D	
base		289.1	250.8	–
V-HPOT		238.8 (-17.4%)	262.6 (+4.7%)	-6.4%

A.3. Impact of the testing data during TTO

In the main paper, we report selecting 5% of data for our test-time optimisation process during experiments, after which we cease weight updates. Table 9 presents performance metrics across varying data quantities. The results demonstrate that our selected 5

However, selecting a percentage of data is not possible in the real-world online scenario, as the data stream's length is unknown, unlike in our experimental case. Thus, we compare the results when using a fixed data quantity for all datasets, equal to 5% of the *H2O* data, which is 960 frames. The observation is a marginal decline in performance.

Our approach, which involves adaptation performed on a relatively small amount of data, only represents a performance trade-off in selecting data quantity and learning rate during test time. We assume that our goal is to achieve the highest performance in the fastest possible time for two reasons: (1) whilst adapting the network faster, more samples result in better predictions, yielding superior final results compared to optimising through the entire test set, and (2) the test-time process has lower computational cost as the optimisation does not process the whole test set for every sample. Due to this, we select a high learning rate ($lr = 0.3$) for V-HPOT during TTO compared to training with a scheduled $lr \in < 0.1, 0.012 >$.

The high learning rate makes the TTO process sensitive to outliers that appear more frequently in larger data subsets and affect vulnerable 3D pose understanding, causing instability and excessive optimisation of network weights and biases. Particularly sensitive is the regression component responsible for understanding depth (z-coordinate), explaining the performance decrease. Using a training-range lr eliminates this effect and maintains performance improvements with 100% data, but increases inference time by $\times 5$ for the test set. Our method and learning rate selection, therefore, balance performance gains with computational efficiency.

A.4. Online vs. Offline TTO

We compare online versus offline adaptation approaches. In the online method, we continuously update weights until processing 5% of the data. Conversely, in the offline approach, we load pre-trained weights for each sample and optimise n times specifically for that sample. Table 10 presents results for the online method alongside offline variants with $n = 1, 3$ and 5. Our findings indicate that the online approach yields superior performance. Additionally, the offline method incurs substantially higher computational costs as n increases.

A.5. Root-based vs. absolute regression

V-HPOT aims to improve absolute depth estimation, which is crucial for specific egocentric applications (e.g., AR/VR and robotics), where metric accuracy is vital for interacting with the physical world. The results demonstrate significant enhancements in absolute 3D pose estimation, while mitigating the root-relative MPJPE-RA error to a lesser extent. Some literature has focused on methods that first estimate hand pose in root-relative space with respect to the wrist joint, e.g., InterHand [7], and then transfer to absolute values. To investigate whether the V-HPOT approach can work by first estimating in root-relative space and then transforming to absolute coordinates, we reimplemented our network using a root-relative training strategy before transferring to absolute values. Table 11 compares the original V-HPOT and the RR-based V-HPOT variant. While the RR variant demonstrates greater improvements in MPJPE-RA, all other metrics decline compared to the original V-HPOT, confirming our design choices.

Table 9. **Ablation study presenting impact** of test-data quantity considered during our TTO process.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
1%	125.9	217.5	67.2	98.5	65.9	344.1	5.7	51.8
5%	53.3	**174.5**	51.1	90.6	**54.1**	253.2	5.8	**36.9**
7%	**52.8**	184.9	51.2	**80.42**	59.3	**185.01**	5.2	41.2
10%	74.8	184.1	**47.1**	84.6	72.2	228.5	**5.1**	40.25
fixed	53.3	175.1	51.1	92.1	**54.1**	252.9	5.8	37.0

A.6. Impact of error in initial pose

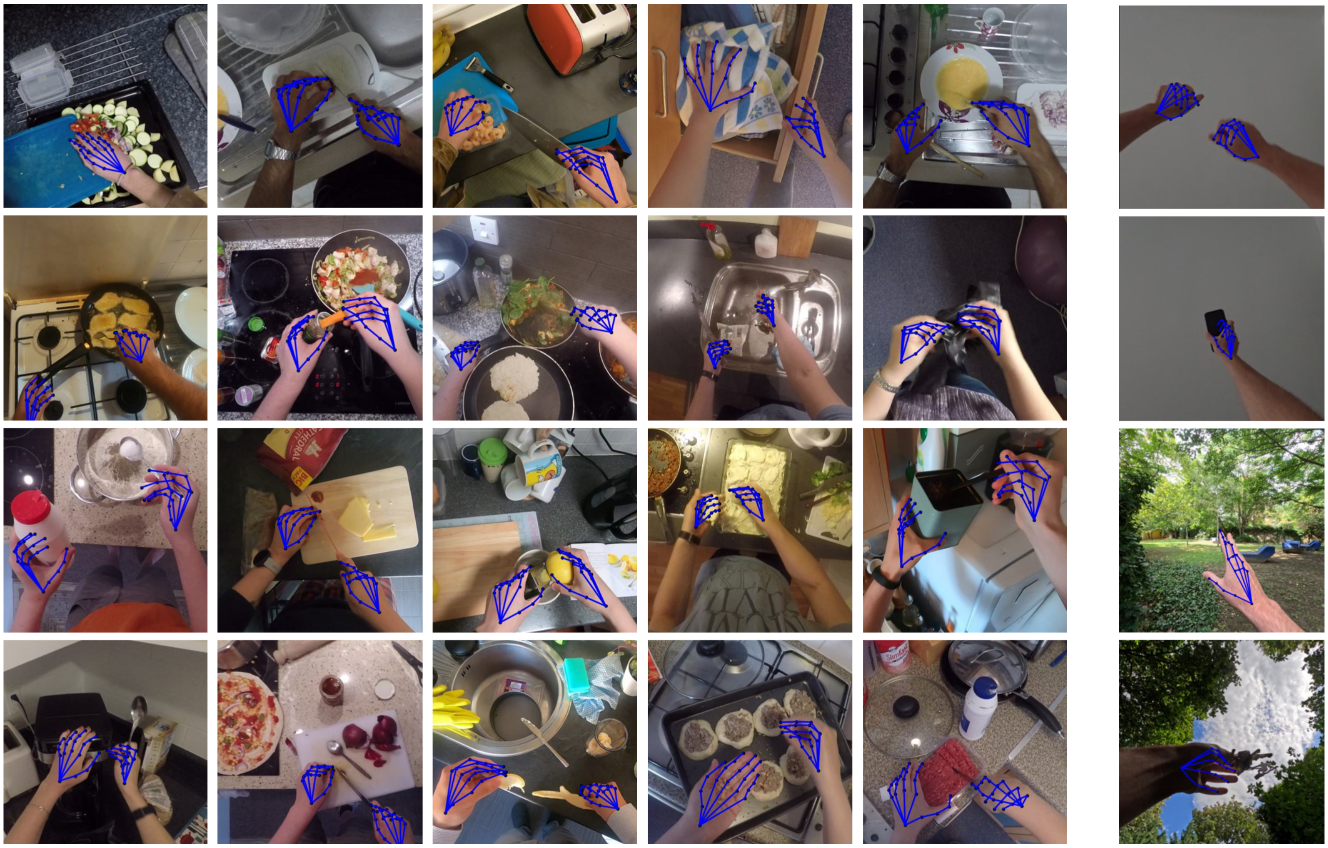
Figure 3 shows the pose losses with and without V-HPOT during testing. In the figure, we can observe that despite a wrong initial prediction, V-HPOT improves the final result in most samples. In this study, we dive deeper into the impact of the initial pose.

To evaluate the robustness of our method to inaccurate initial predictions, we conduct a systematic noise augmentation experiment. We add uniform noise sampled from $< \frac{-n}{2}, \frac{n}{2} >$ millimetres to the initial 3D hand poses prediction, where $\in \{10, 20, 35, 40, 50\}$. This simulates scenarios where the initial pose estimation contains higher error (MPJPE + noise), testing whether our TTO approach can recover from poor initialisations or if it reinforces incorrect predictions.

Results in Table 12 demonstrate that our approach exhibits degradation rather than catastrophic failure. Higher degradation is observed in *H2O*, as the initial predictions are more accurate than in *AssemblyHands*. The error is higher in the *AssemblyHands* initial prediction, and noise has a lesser impact on it. Even with maximum noise $n = 50$, we observe improvements over *base* (not using V-HPOT).

A.7. Qualitative results in Epic-Kpts

Datasets for egocentric 3D hand pose are limited to a laboratory environment due to annotation methods relying either on a multi-view camera setup or motion capture technology. This limits the recording environments, as all data is confined to indoor laboratory scenes. In contrast, hand pose estimation aims to work in an in-the-wild setup with a more diverse environment. We show qualitative results of our method on the real-world dataset *Epic-Kpts* [11], which



(a) Epick-Kpts

(b) Custom images

Figure 5. **Qualitative results of our method on the in-the-wild *Epic-Kpts* and own images.** The estimated 3D pose is projected into 2D space, as no ground truth 3D pose labels are available. Despite the challenging natural environments full of everyday objects, various backgrounds, and sceneries, V-HPOT produces geometrically accurate poses. In *Epic-Kpts*, camera wearers perform dynamic actions involving manipulating various objects while preparing meals or cleaning. In addition to *Epic-Kpts*, we capture images without depth cues, featuring an outdoor background and a subject pointing at the wall, without any reference points.

Table 10. **Ablation study presenting impact** of online and offline TTO process strategy selection.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
$n = 1$	153.3	306.2	66.3	105.3	74.3	277.7	6.4	96.9
$n = 3$	154.4	305.7	66.7	105.3	75.0	278.4	6.3	96.6
$n = 5$	154.1	305.9	66.8	105.3	75.1	278.6	6.4	96.6
Online	**53.3**	**174.5**	**51.1**	**90.6**	**54.1**	**253.2**	**5.8**	**36.9**

is a subset of the *Epic-Kitchens* [4] dataset. Captured in the various kitchens where subjects prepare meals, it is full of natural actions, motion blur, and a wide variety of interacting objects. Unfortunately, there are no 3D pose labels for comparison, limiting us to qualitative analysis only. In addition to Epic-Kpts, we capture images with different background structures where hands are in front of a wall without any reference information, and outdoors in the gar-

Table 11. **Root-based vs. absolute regression.** Comparison of V-HPOT with regression of absolute coordinates against approach where regressions is done relative to wrist and then transformed to absolute coordinate system.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
Root-based approach								
base	133.5	299.7	58.1	102.5	64.2	268.7	6.6	87.1
V-HPOT	57.1	209.1	**40.5**	91.7	57.2	278.5	6.3	47.2
Δ	57.3%	30.2%	30.2%	10.5%	10.9%	3.7%	4.2%	45.8%
Absolout depth approach								
base	179.6	297.7	52.7	105.3	126.2	301.8	7.4	80.1
V-HPOT	**53.3**	**174.5**	51.1	**90.6**	**54.1**	**253.2**	**5.8**	**36.9**
Δ	70.3%	41.2%	3.0%	12.5%	57.2%	16.2%	21.8%	53.9%

Table 12. **Impact of error in initial pose prediction** on V-HPOT performance. Results shown for different noise levels (in mm) applied at TTO.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
base	179.6	297.7	52.7	105.3	126.2	301.8	7.4	80.1
V-HPOT	53.3	174.5	51.1	90.6	54.1	253.2	5.8	36.9
n=10mm	56.3	185.8	51.6	92.3	53.0	229.0	5.6	44.1
n=20mm	71.5	185.4	54.8	92.3	52.0	228.3	5.6	44.0
n=30mm	94.8	190.8	58.6	90.6	53.5	225.5	5.7	46.0
n=40mm	122.5	201.5	62.9	89.5	60.6	223.3	5.7	48.7
n=50mm	149.2	211.2	67.2	88.4	69.1	218.0	5.7	50.3

den, where the background is far from the camera. Results in Figure 5 show that our V-HPOT result in accurate pose estimates even in such challenging scenes.

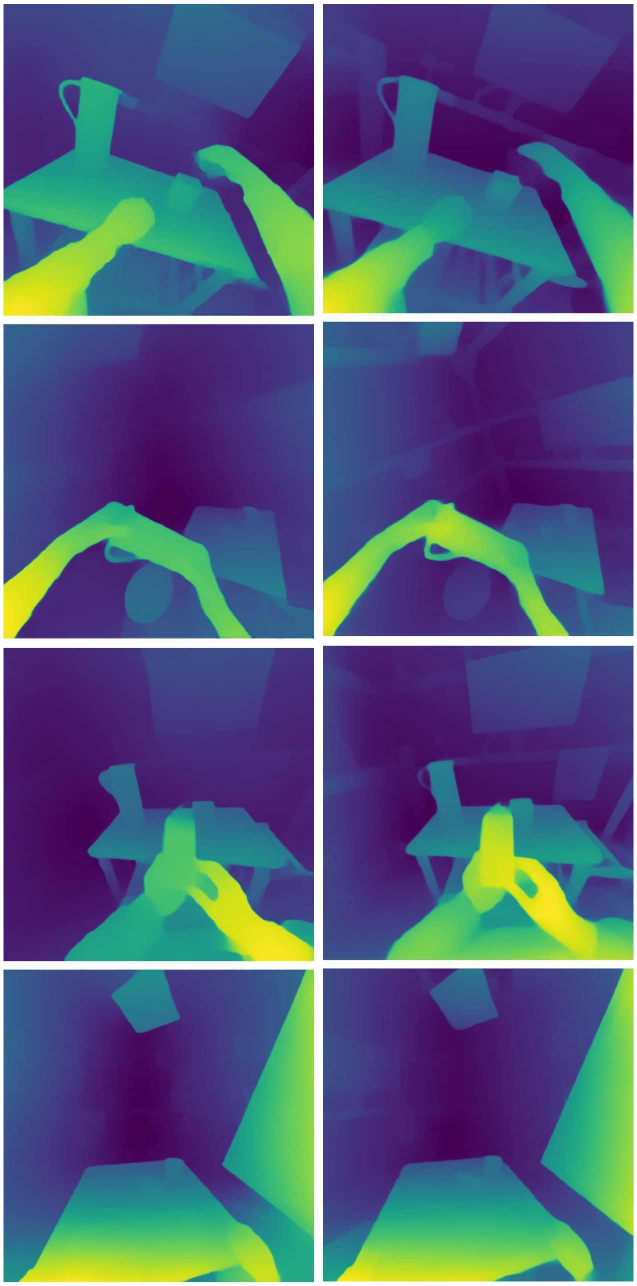
A.8. Qualitative analysis of pseudo-depth estimation models

In the main paper, we compare two pseudo-depth estimation models: *DPT-Hybrid* [12] and *DepthAnything* [16]. The results vary depending on the metric used. We measured inference using an NVIDIA RTX 3090 GPU over 1000 trials. Given the mixed quantitative results, *DPT-Hybrid* is faster (26.8 ms vs. 39.8 ms), which is essential at the experimental stage. Furthermore, we visually analyse the differences between *DPT-Hybrid* and *DepthAnything* by imagining random frames from the *HOT3D* dataset. These frames are presented in Figure 6. We observe that both models perform similarly in terms of hand regions. The main difference is *DepthAnything*'s superior understanding of the background details. In our study, which focuses solely on the hands, these background details are less important, while the faster inference and better absolute 3D error in pose estimation favour *DPT-Hybrid* for the role of our auxiliary task.

A.9. Ground truth depth vs. pseudo-depth analysis

Pseudo-depth estimation, used as an auxiliary task, is limited by errors that can propagate during hand pose estimation training. Among the datasets in this study, only *H2O* provides sensor-based ground-truth depth measurements, limiting the possible cross-domain experiments. To evaluate the impact of depth estimation quality, we train our network on the *H2O* dataset and assess its performance in a cross-domain scenario on *HOT3D* and *AssemblyHands*, using ground-truth depth and *DPT-Hybrid* pseudo-depth as auxiliary tasks.

Results reveal a negligible difference for *base* (without TTO) (0.3% MPJPE) between GT depth versus pseudo-depth when averaged across both domains. With TTO applied, the results are mixed. For *HOT3D*, the difference in MPJPE is minimal in favour of the pseudo-depth (262.6 vs. 263.2), while for AssemblyHands is larger (238.8 vs 250.7). This suggests that pseudo-depth estimation provides sufficient spatial correctness to serve as an auxiliary task. In some cases, it even improves performance through a regularisation effect. The less detailed background informa-



(a) DPT-Hybrid (b) DepthAnything

Figure 6. A visual comparison of the *DPT-Hybrid* and *DepthAnything* pseudo-depth estimation models. *DepthAnything* is better at estimating background details, while the quality of hand estimates is similar in both methods.

tion in slightly imperfect pseudo-depth may prevent overfitting to specific depth patterns coming from the scene background or sensor noise, but being not crucial for hand understanding, leading to better generalisation across domains and serving effectively as an auxiliary task. Additionally, pseudo-depth remains advantageous as an auxiliary task because it can be used on any dataset without requiring a depth sensor.

References

[1] Ahmed Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. THOR-Net: End-to-end Graformer-based Realistic Two Hands and Object Reconstruction with Self-supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1001–1010, 2023. 1

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7061–7071, 2025. 1

[3] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-Based Unified Recognition of Two Hands Manipulating Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4769–4778, 2023. 1

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[5] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 571–580, 2020. 1

[6] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two Hands Manipulating Objects for First Person Interaction Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 1

[7] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020. 2

[8] Wiktor Mucha, Michael Wray, and Martin Kampel. SHARP: Segmentation of Hands and Arms by Range Using Pseudo-depth for Enhanced Egocentric 3D Hand Pose Estimation and Action Recognition. In *International Conference on Pattern Recognition*, pages 178–193. Springer, 2025. 1

[9] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem 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

[10] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing Hands in 3D with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 1

[11] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3D Hand Pose Estimation in Everyday Egocentric Images. In *European Conference on Computer Vision*, pages 183–202. Springer, 2025. 1, 2

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 4

[13] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 1

[14] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified Egocentric Recognition of 3D Hand-object Poses and Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019. 1

[15] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21243–21253, 2023. 1

[16] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 4