Predicting 4D Hand Trajectory from Monocular Videos

Yufei Ye[1] Yao Feng[2] Omid Taheri[2] Haiwen Feng[2]
Shubham Tulsiani[1*] Michael J. Black[2*]
[1]Carnegie Mellon University [2] Max Planck Institute for Intelligent Systems

https://judyye.github.io/haptic-www

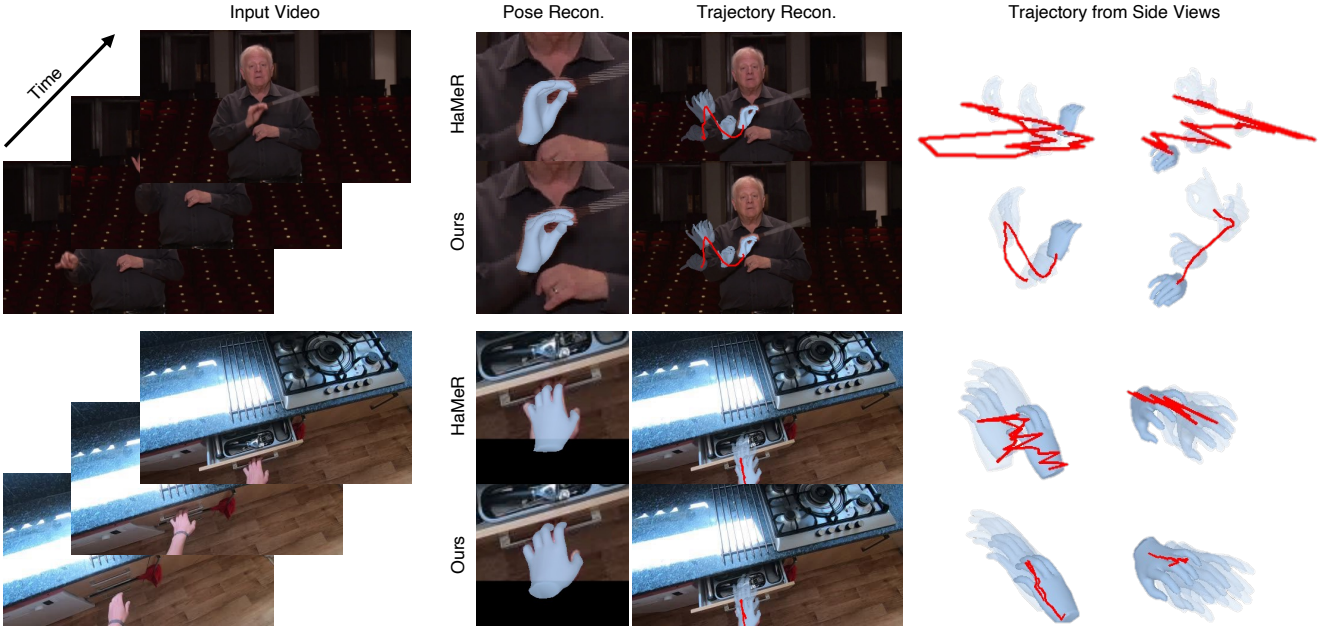

Figure 1. Given a monocular video depicting hand motion, HaPTIC reconstructs both **Ha**nd **P**ose (Pose Recon.) and 4D hand **T**rajectory **i**n consistent global **C**oordinate (last 3 columns). The existing method produces convincing reprojection but its 4D trajectory is not plausible (side view). In contrast, our method can generate faithful 4D trajectories. The opaque hand shows reconstruction of the last frame while semi-transparent hands visualize reconstructions from previous frames. Red curves visualize root trajectories.

Abstract

We present HaPTIC, an approach that infers coherent 4D hand trajectories from monocular videos. Current video-based hand pose reconstruction methods primarily focus on improving frame-wise 3D pose using adjacent frames rather than studying consistent 4D hand trajectories in space. Despite the additional temporal cues, they generally underperform compared to image-based methods due to the scarcity of annotated video data. To address these issues, we repurpose a state-of-the-art image-based transformer to take in multiple frames and directly predict a coherent trajectory. We introduce two types of lightweight attention layers: cross-view self-attention to fuse temporal information, and global cross-attention to bring in larger spatial context. Our method infers 4D hand trajectories similar to the ground truth while maintaining strong 2D reprojection alignment. We apply the method to both egocentric and allocentric videos. It significantly outperforms existing methods in global trajectory accuracy while being comparable to the state-of-the-art in single-image pose estimation.

1. Introduction

Consider the video frames shown in Figure 1. One can infer not only the articulation of the hand (3D pose) depicted in every frame but also understand the hand motions across frames and in space. For example, in the bottom row of Figure 1, we understand that the hand extends towards the drawer and then retracts along a nearly identical path. The ability to infer 4D hand trajectory (3D space plus time) is important for many downstream tasks, such as reasoning about hand-object interactions in a global scene [40], AR/VR applications [47], and imitation learning in

* Equal contribution.

robotics [1, 61]. Unfortunately, despite impressive progress in hand pose estimation that infers frame-wise 3D pose, current methods still struggle to coherently put the hands into a global 3D space. In this work, we address this problem with a system that can infer coherent 4D hand trajectory and pose from monocular videos.

While no prior work is dedicated to direct 4D hand trajectory prediction, common practices often involve first predicting per-frame 3D hand poses, 'lifting' them to a world coordinate system, followed by test-time optimization [9, 22, 44]. The de-facto lift method uses a weak-to-full perspective transformation (Weak2Full) [46, 51], which places the predicted hand at a certain distance given the camera intrinsics and predicted scale (Sec. 3.1). However, we find that this operation introduces a large error in 4D motion, and that even post-processing optimization struggles to correct it. An alternative lift approach uses estimated depth [2, 60] but occlusion of the hand from object or scene interaction, or from the other hand, makes the induced trajectory inaccurate. One alternative philosophy is to consider the more holistic task of full body estimation in 4D and then the hands are defined relative to the body [54, 69]. Unfortunately, this requires the full body to be largely visible which is typically not true for videos focusing on hand manipulation and egocentric video.

In contrast, we formulate the problem as a 4D inference problem from video and leverage an implicit data-driven prior. Given a video sequence as input, we output MANO [50] hand parameters with the wrist in global coordinates. Unlike hand-designed lifting, our approach better captures prior information about hand motion and does not rely on explicit full-body estimation. Our approach, along with the concurrent works of [72, 74], is among the first to present a feed-forward method for estimating consistent 4D hand trajectories from monocular video.

The problem, however, is the lack of training data containing video paired with 3D hand annotations in global coordinates. While some such data exists, it is much scarcer and less diverse than image-based training data. To address this limitation, our key idea is to heavily leverage the single-image data in terms of both models and training data. We then incorporate video training data in such a way that we need significantly less data than if we trained the 4D inference with video alone. Specifically, we repurpose the state-of-the-art image-based hand pose estimator [46] to incorporate video inference by injecting two types of lightweight attention layers (Fig. 2). The first is a cross-view self-attention layer that sees across multiple frames to leverage temporal cues. The second is a global-context cross-attention layer that sees the original frames (from which the input hand images are cropped) to gain a larger field-of-view of the scene. By design, the network allows both video and images as input — when the input is an image, the cross-view attention simply degrades to attending to itself. This allows us to intersperse both small-scale video data and large-scale image data in training batches. In this way, our new model, called HaPTIC, maintains the robustness and generalization ability of single-image methods but enriches them with temporal information from video.

We evaluate HaPTIC on both allocentric and egocentric videos. We analyze the drawbacks of the prevalent practices, including when and why they fail. HaPTIC significantly outperforms all other potential candidates in terms of global trajectory accuracy. It also provides a better initialization for test-time optimization. We carefully ablate our method to analyze our design choices. When HaPTIC serves as an image-based pose estimator, we find that it even outperforms the state-of-the-art image-based method in terms of per-frame hand pose 2D alignment. Finally, we show the generalization ability to in-the-wild videos and images by presenting more qualitative results.

2. Related Work

Image-Based Hand Pose Estimation. There is a rich body of work tackling hand pose estimation from image input and typical approaches can be categorized into template-based methods [51, 75] and template-free methods [32, 35, 36, 41, 59, 76]. Both benefit from diverse large-scale annotated data [26, 42, 46, 64, 78] and foundational backbones [8, 20, 23]. While these methods perform effectively on single image frames, their predictions are all in the canonical local frame of the hand. Their predictions can be reprojected back to each video frame but do not induce coherent 4D trajectories in space. Yet, we believe that these models can establish a solid foundation for video-based hand estimation. Specifically, our work is built upon the state-of-the-art image model, HaMeR [46], which is a large transformer trained on large-scale datasets to predict the parameters of a 3D hand model (MANO [50]).

Video-Based Hand Pose Estimation. Hand tracking in 3D has a long history and is central to many applications. To get the most reliable tracking, methods rely on specialized hardware like IMUs or magnetic sensors [5, 10, 15, 63]. Multi-camera video systems can also provide pseudo ground truth [11, 30, 42, 62]. While recent approaches reconstruct hand motion from monocular videos [22, 37, 38, 44, 68, 72], these methods are based on test-time optimization using a learned prior, which is time-consuming; they can take a couple of minutes to hours to process clip of a few seconds. There are a few feed-forward video-based methods [11, 13, 29, 34, 65] but these are not as robust or general as the single-image methods due to a lack of annotated video training data. More importantly, they focus on improving local pose and do not tackle the problem of global trajectory estimation. Dyn-HaMeR[72] and HaWoR[74] are two concurrent works that both recov-

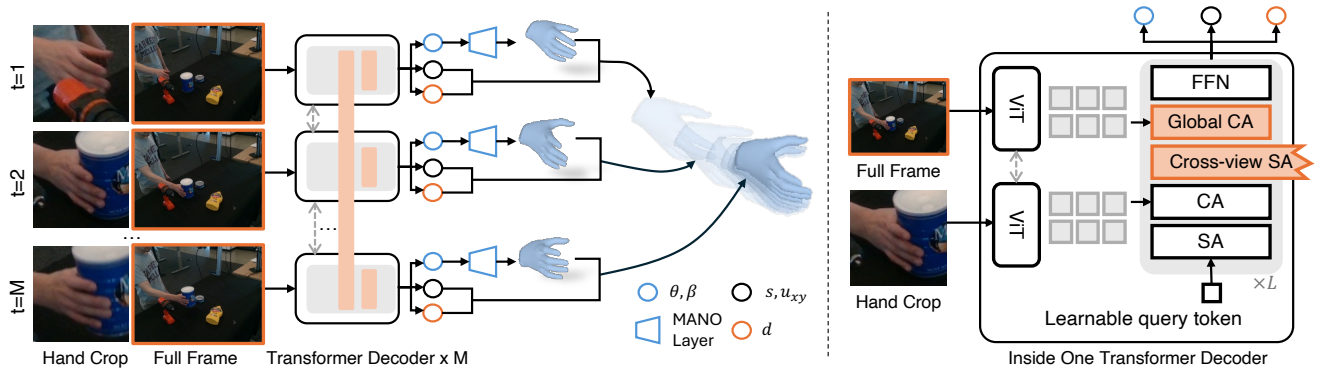

Figure 2. **Overall pipeline (left)**: HaPTIC extends image-based model HaMeR. HaPTIC takes in M frames at a time and passes them through image towers that share weights. Each image tower outputs MANO parameters in local coordinate, and trajectory parameters d, u_{xy} that directly places the predicted local hand to global 4D trajectory. **Inside one image tower (right)**: The image tower is based on transformer decoder. For each block, we add a cross-view self-attention layer (Cross-view SA) to fuse temporal information from other frames and a cross-attention (Global CA) to features of the original frames. Orange indicates new components introduced by ours compared to HaMeR.

ers global pose. Dyn-HaMeR is an optimization approach that takes 10 min for an 128 frames while ours only takes 4s. HaWoR explores genrative motion prior to infer motion sequence. Our approach is a feed-forward method that can behave as both image and video model, with its architecture specifically tailored to the limited availability of video data.

4D Whole-Body Reconstruction. 4D whole-body trajectory estimation in global space has seen recent attention, either using feed-forward predictions [31, 45, 54, 58, 71, 73], or global optimization [16, 17, 57, 67] that decouples human motion from camera motion [52] using learned human motion prior [48]. These methods could potentially place hands in global coordinates by explicitly attaching the hand pose to the wrist. However, this requires the full body to be largely visible, which is often not the case in videos focusing on hand manipulation and egocentric videos. Instead of leveraging whole body pose as a form of explicit global context, the global spatial cross-attention in HaPTIC provides the implicit context of the scenes and humans. We show that even when the full body is visible, our method outperforms the state-of-the-art method [54] in this category.

Adapting Image-Based Models to Videos. We are inspired by the idea in generative AI that upgrades pretrained single-image models [7, 28, 49] to other modalities through lightweight adaptation. Image-based models are upgraded to video generative models via multi-frame hierarchical training [24], spatiotemporal factorization [56], video noise priors [18], and temporal attention combined with 3D convolutional blocks [3, 4]. Image models are similarly extended to 3D generation [14, 25, 39, 53] by introducing multi-view consistency or geometry-guided attention mechanism. In our work, we repurpose a pretrained single-image hand estimator to take in multiple frames. Importantly, this allows us to train on both video and image datasets.

3. Method

Given a monocular video with detected hand bounding boxes $\{I_t, B_t\}_{t=1}^{T}$, HaPTIC infers hand articulations and their 3D locations in the camera coordinate system in a feed-forward manner. We first revisit the preliminary parametric hand model MANO and analyze why the de-facto approach for uplifting is undesirable (Sec. 3.1). We then introduce our simple yet effective parameterization of hand trajectory (Sec. 3.2). Finally, we describe our network architecture that predicts hand trajectories and explain our design choices, which address the problem of limited video training data (Sec. 3.3 and 3.4).

3.1. Preliminaries: MANO and Weak2Full

Like prior work [13, 46, 51], we use the parametric hand model, MANO [50]. MANO takes a 48-dim pose parameter vector $\boldsymbol{\theta}$, including finger articulation and wrist orientation, as well as 10-dim shape parameter vector $\boldsymbol{\beta}$ and it outputs a hand mesh surface $\mathcal{M}^l(\boldsymbol{\theta}, \boldsymbol{\beta})$ and hand joints $\boldsymbol{J}^l(\boldsymbol{\theta}, \boldsymbol{\beta})$. Superscripts, l, denote coordinates in the local MANO frame.

Weak-to-Full Perspective. In order to overlay the predicted (local) hand mesh onto the input image crop, prior work also predicts a weak-perspective camera, which projects a 3D point by scaling and translating the xy components, $\boldsymbol{X}^i_{x,y} = s\boldsymbol{X}^l_{x,y} + u_{x,y}$, where superscript i denotes coordinate in the image frame. Note that although this results in 2D image alignment, a weak-perspective camera does not give the 3D position in the camera space $\boldsymbol{X}^c$. To obtain a 3D position, the widely adopted approach is to use weak-to-full perspective transformation (Weak2Full) [19, 22, 27, 33, 46, 51]. It uses an assumed focal length f as input and places the hand at a distance such that the re-projection of the transformed hand is approximately at the predicted scale under that assumed per-

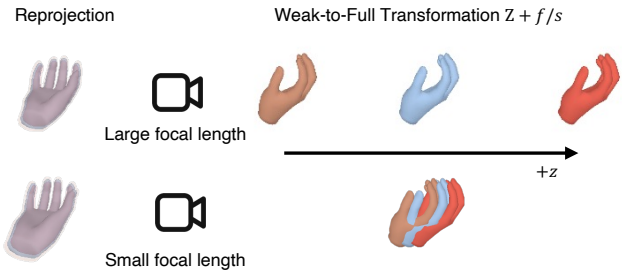

Figure 3. **Toy example of Weak2Full transformation:** The predicted scale s of both yellow and red hands are only 6% off the blue hand. Their reprojections appear similar yet the 3D position induced by Weak2Full varies a lot with large focal length.

spective cameras $\boldsymbol{X}^c = \boldsymbol{X}^l + (u_x, u_y, f/s)$.

What's wrong with Weak-to-Full? Looking closely at the offset in the Z-axis, f/s, in the Weak2Full process, we notice that the error in the predicted scale s affects the 3D location (4D trajectory for videos). As shown in Fig. 3, the trajectory is more prone to error in the predicted scale when the focal length is larger. Furthermore, even with optimal scales, different estimates of the focal length induce different motions in 3D space.

3.2. Parameterize 4D Hand Trajectory

In contrast with the Weak2Full transformation that parameterizes hand trajectory with scale and focal length, *i.e.* Weak2Full($\{u_{x,t}, u_{y,t}, s_t, f_t\}_{t=1}^T$), we advocate for directly representing hand trajectory in the metric space, which is actually the most natural motion parameterization. Specifically, we simply predict *change of depth* relative to the first frame: $\boldsymbol{X}^c = \boldsymbol{X}^l + (u_x, u_y, \Delta d + d_1)$. d_1 is the estimated offset from the first frame, either from the Weak2Full transformation on the first frame or from depth prediction. We predict metric depth *change*, which is invariant to the distance to the camera, instead of absolute metric depth because of depth ambiguity – it is hard to tell if the hands are $5m$ away or $5.5m$ away but easier to infer that hand is moving by $5cm$ between frames. We learn a network to predict these parameters $\Delta d_t, u_{xy,t}$ along with $\boldsymbol{\theta}_t$ directly from video input.

3.3. Repurpose Image-Based Model for Videos

Training a generalized video model from scratch requires a lot of data. Due to the lack of 3D annotated video data, we choose to extend the state-of-the-art hand pose image-based model (HaMeR) and design our network architecture to best leverage the existing model. We then finetune its weights to repurpose it into a video model.

Multi-Frame Siamese Network. HaPTIC, like HaMeR, is also a transformer-based model. Instead of taking a hand crop from a single image, we modify the network to take in M crops at a time, along with their full frames (Fig. 2 left). Each frame is processed in a Siamese-network manner (sharing weights). The network is adapted from HaMeR with two types of lightweight adaption layers to fuse information among frames. In addition to the original prediction heads to output MANO parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$ and weak-perspective cameras parameters $s, u_{x,y}$, we introduce an additional prediction head to predict metric depth change in order to uplift local hands to global hand trajectory.

Light-Weight Adaption Layers. Each frame (and crop) is first passed to a ViT backbone, followed by a transformer decoder. As shown in Fig. 2 right, inside each block of the transformer decoder, we insert a cross-view self-attention (SA) layer and a global cross-attention (CA) layer after the original CA layer. The cross-view SA layer fuses temporal information from other frames. We add a positional encoding of the frame number in this layer. Since hand crops alone discard global information, the global cross-attention layer takes in a larger-view context by attending to the full-frame ViT features. The overall architecture only adds 4% new parameters. We initialize the new attention layers to output zero such that the new layers do not destroy the pre-trained structures at the start of training.

Generalize to Image Reconstruction. While HaPTIC is designed for video reconstruction, it can naturally incorporate single images as input without any modification. In such a case, cross-view SA only attends to its own frame. The image-mode inference can simply be achieved by a reshape operation.

3.4. Learn with Interspersed Video and Image Data

We believe that generalization ability of video models can be achieved by training on diverse *image* datasets with only limited video data. For each training batch, half of the batch comes from video datasets and is trained in video mode, while the other half comes from image datasets and is trained in image inference mode.

Losses. We leverage the best practices for parametric hand pose estimation from images and supervise the model with a combination of 4D (video), 3D, and 2D losses.

We supervise the predicted hand trajectories of length M in a global space with the ground truth joints in global coordinates: $\mathcal{L}_{\text{4D}} = \sum_{t=1}^{M} \|\boldsymbol{J}_t^c - \hat{\boldsymbol{J}}_t^c\|_1$. Note that this global supervision sounds demanding but global coordinates are actually present in most multi-view capture datasets. We align joint trajectories by the root joint in the initial frame.

The rest of the losses follow the recipe of training image-based hand pose reconstruction. We directly supervise hand pose in the local frame in 3D space, including consistency of MANO parameters and 3D hand joints: $\mathcal{L}_{\text{3D}} = \|\boldsymbol{J}^l - \hat{\boldsymbol{J}}^l\|_1 + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$. We also supervise the network with 2D keypoint reprojection loss: $\mathcal{L}_{\text{2D}} = \|\boldsymbol{j} - \hat{\boldsymbol{j}}\|_1$. Lastly, 2D keypoint consistency alone may lead to unrealistic 3D

Local Pose	Uplift	ARCTIC-EXO			DexYCB			ARCTIC-EGO		
		GA-MPJPE	FA-MPJPE	ACC-NORM	GA-MPJPE	FA-MPJPE	ACC-NORM	GA-MPJPE	FA-MPJPE	ACC-NORM
HaMeR	Weak2Full	7.88	21.79	81.19	3.28	12.66	14.39	2.91	5.53	11.95
HaMeR†	Weak2Full	5.53	13.29	30.09	1.93	9.78	4.42	3.48	9.23	9.62
HaMeR†	ZoeDepth	3.28	10.09	1.83	4.93	20.51	1.32	2.95	5.60	1.57
HaMeR†	DepthAnythingV2	3.37	10.99	1.84	4.95	20.59	1.32	5.45	9.32	1.63
HaMeR†	WHAM	4.79	14.41	2.43	–	–	–	–	–	–
ArcticNet-LSTM	Weak2Full	7.41	16.78	6.06	–	–	–	2.90	4.61	5.24
HaPTIC (Ours)		1.88	4.49	1.48	1.80	5.80	1.70	2.39	4.49	1.34

Table 1. **Comparison with baselines.** We compare our method with different uplifting baselines on two allocentric datasets (ARCTIC-EXO and DexYCB) and one egocentric dataset (ARCTIC-EGO). We report trajectory errors in GA-MPJPE, FA-MPJPE, and ACC-NORM. † denotes finetuned models on top of the publicly released HaMeR. Among them, dark yellow marks the best results and light yellow marks the 2nd best ones.

poses. To encourage the generated hands to look realistic, we use an adversarial loss following prior work [19, 46]: $\mathcal{L}_{adv} = (D(\boldsymbol{\theta}, \boldsymbol{\beta}) - 1)^2$.

4. Experiments

We train our model a combination of 5 video datasets and 10 image datasets and evaluate the performance of the our model in terms of 4D trajectory (Sec. 4.1) and 3D poses (Sec. 4.2). We evaluate our model on two allocentric datasets [6, 11] and one egocentric dataset [11]. In Sec. 4.3, we analyze the effect of our network design. Qualitative results on in-the-wild results from both videos and images are further shown in Sec. 4.4 and the **Suplemental Video**.

Training Data and Setup. We train and evaluate our model on the following datasets:

- DexYCB [6] is an allocentric video dataset captured by multiple calibrated cameras. It features humans handling rigid objects on cluttered tables. Each sequence is 2-3 seconds long. We use the standard s0 split.
- ARCTIC [11] provides both allocentric (-EXO) and head-mounted (-EGO) views of bi-manual manipulation of articulated objects. Each sequence is ∼30 sec. We evaluate on the official subject-wise test split.
- Combination of image data: Our auxiliary image dataset is the same as the HaMeR training set, which consists of 10 datasets [6, 12, 21, 26, 30, 42, 55, 64, 77, 78].
- Other video data for training: This includes the 4 datasets from the auxiliary image dataset that have 3D world coordinate annotations [6, 21, 30, 42], as well as the ARCTIC dataset. The previous four are all allocentric, while only ARCTIC also contains egocentric views.

We set the winow size of our model to 8, *i.e.* $M = 8$. During training, we augment the frame rate up to 6 fps to accommodate different speeds of hand motion. At test time, we use a sliding window with a 1-frame overlap to process long videos. More implementation details are in *Sup. Mat.*

Baselines. While there is a plethora of work to estimate per-frame (local) hand poses, there are no direct baselines that reconstruct 4D hand trajectories in a feed-forward manner. Therefore, we compare with common practices and possible alternatives to uplift local poses to 4D. The local hand poses for all baselines are from the same state-of-the-art image-based hand pose estimator [46] since local pose is not the main focus of our paper. For a fair comparison, we also finetune the released HaMeR further, denoted as †.

- We first compare with the most commonly used uplift method, Weak2Full, with unknown camera intrinsics, which we set to the diagonal length of the image [33, 54].
- Metric depth computed using a monocular depth estimator can also be used to uplift local hand pose. We use ZoeDepth [2] or DepthAnythingV2 [66] to predict pixel-wise metric depth map and solve for the optimal 3D offset that align it with the predicted hand surface from the local hand pose.
- WHAM [54] estimates body pose in global coordinates without hand pose. We use WHAM's estimated wrist orientation and location to uplift estimated local hand articulation. Since WHAM requires most of the body to be visible, this method only applies to ARCTIC-EXO.

Evaluation Metrics. We use metrics from whole-body pose estimation [54, 70, 73] to evaluate 4D pose trajectory. In contrast to aligning poses per frame, GA-MPJPE globally aligns hand joints at all frames within one sequence before calculating its Mean Per Joint Position Error (MPJPE). FA-MPJPE aligns the whole hand pose trajectory only using the pose in the first frame. ACC-NORM computes acceleration error compared with ground truth for each joint. The first two metrics evaluate trajectory quality globally while ACC-NORM focuses on local trajectory from adjacent frames. See *Sup. Mat.* for evaluation details.

4.1. 4D Trajectory Reconstruction

We report quantitative comparisons on three datasets in Table 1 and visualize the results in Fig. 4.

While finetuning the general HaMeR model further improves its own performance, the Weak2Full transformation produces significant jitter in global space. This happens even when ground truth intrinsics are given. Our results highlight that trajectories from Weak2Full are sensi-

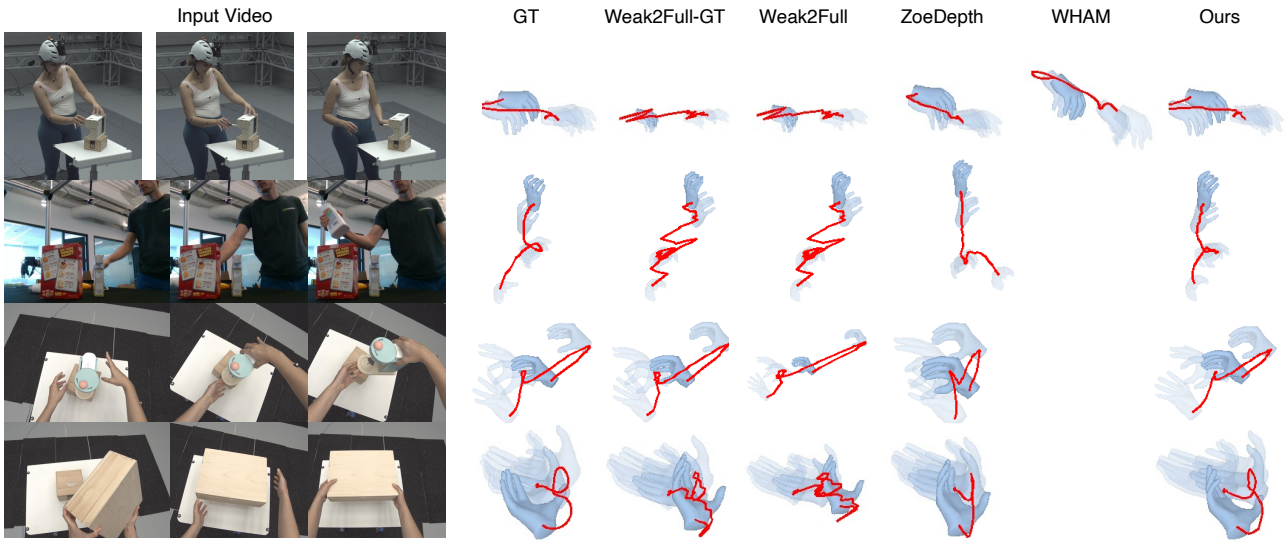

Figure 4. **Qualitative comparison.** We compare our approach qualitatively with other feed-forward baselines on all of three datasets. We show the first, middle, and last frames of an input video. We visualize hand root trajectory as red curves and five poses along time in a global coordinate viewed from the side. Poses from previous frames are visualized with transparency while the last frame reconstructions are opaque. We only visualize the right hand for clarity but all methods can reconstruct both the left and right hand. We encourage readers to see results in videos on our website.

tive to camera intrinsics. The error becomes even bigger with a large focal length (ARCTIC-EXO). Weak2Full only performs well in terms of global alignment of trajectory (GA/FA-MPJPE) under the egocentric setting with ground truth intrinsics after finetuning. This is because the egocentric camera has a smaller focal length and, when hands are closer to the camera, it is reasonable to estimate hand distance by using its scale. However, when the intrinsics deviate from ground truth, the trajectory induced from Weak2Full degrades.

Uplifting by metric depth prediction gives a smoother trajectory (lower acceleration error). However, it does not perform well with occlusions, which are present in the more cluttered scenes in DexYCB or due to hand-object interactions on ARCTIC-EGO. With better metric depth prediction [66], interestingly, we find limited improvement in exocentric view and, surprisingly, even some degradation in egocentric view. This is probably because recent progress in depth estimation mostly focuses on sharper edges on natural or indoor images.

Global human body poses from WHAM only work when the person is largely visible, thus this approach cannot be used on DexYCB or egocentric videos. On ARCTIC-EXO, it still does not perform well because the method is optimized to predict the body's root trajectory instead of trajectories of distal extremities.

In contrast, while HaPTIC is able to predict equally good 2D alignment (Fig. 1, 6 and *Sup. Mat.*), Fig. 4 shows that our global trajectories are more consistent with ground truth. Quantitatively, HaPTIC is among the top two methods across all metrics on all datasets.

	GA-MPJPE	FA-MPJPE	ACC-NORM	PA-MPJPE
Weak2Full-GT + Opt.	4.38	10.90	1.14	0.54
ZoeDepth + Opt.	3.75	9.16	0.92	0.68
HaPTIC (Ours) + Opt.	**2.28**	**5.69**	**0.90**	**0.51**

Table 2. **Comparison with test-time optimization.** We compare trajectories after optimization when they are initialized from different baselines on a subsplit (0th view) of ARCTIC-EXO.

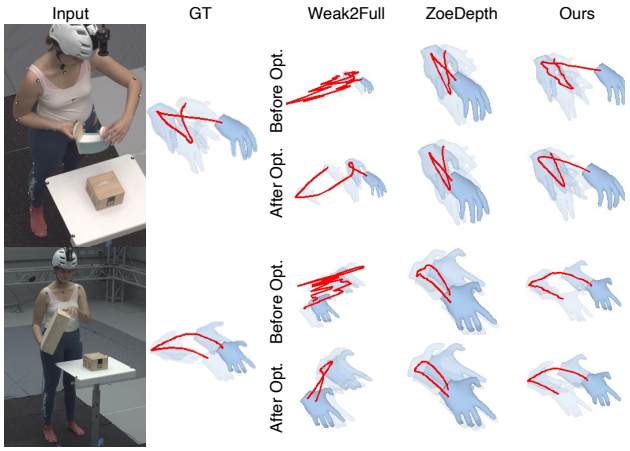

Figure 5. **Qualitative comparison of optimization**: We compare results before and after optimization with trajectories initialized from baselines and HaPTIC. The optimized trajectories are smoother but optimization struggles to correct global error.

Relation to Concurrent Work. HaWoR [74] is a concurrent work that learns a generative prior to directly predict hand motion in the world space. We predict in the camera 3D space, which is same with the world space when cam-

Method	New Days			VISOR		
	@0.05	@0.1	@0.15	@0.05	@0.1	@0.15
FrankMocap [51]	16.1	41.4	60.2	16.8	45.6	66.2
METRO [36]	14.7	38.8	57.3	16.8	45.4	65.7
MeshGraphormer [35]	16.8	42.0	59.7	19.1	48.5	67.4
HandOccNet (param)[43]	9.1	28.4	47.8	8.1	27.7	49.3
HandOccNet (no param) [43]	13.7	39.1	59.3	12.4	38.7	61.8
HaMeR [46]	48.0	78.0	88.8	43.0	76.9	89.3
HaPTIC (Ours)	48.6	79.0	89.9	46.8	79.5	90.9

Table 3. **Evaluation of single image pose.** We report PCK at different thresholds on benchmark HInt. We compare HaPTIC with multiple image-based hand pose estimation baselines.

Method	ARCTIC-EXO		DexYCB		ARCTIC-EGO	
	PA-MPJPE	RA-MPJPE	PA-MPJPE	RA-MPJPE	PA-MPJPE	RA-MPJPE
ArcticNet-LSTM	1.03	2.29	–	–	1.04	2.38
Deformer	–	–	0.52	1.36	–	–
HaMeR	0.63	1.34	0.42	0.92	0.71	1.54
HaPTIC (ours)	0.62	1.40	0.44	0.93	0.72	1.52

Table 4. **Evaluation of single image pose.** We report PA/RA-MPJPEG where poses are aligned *per-frame*.

eras are static. On DexYCB, their GA/FA-MPJPE is 2.98 and 9.38, respectively. It suggests HaPTIC produces better hand motion in allocentric views or under static cameras.

Test-Time Optimization. Can test-time optimization improve jittery trajectories from feed-forward methods? We further optimize the predicted trajectories for 1000 iterations on a subset (view 0) of ARCTIC with respect to 1) reprojection error to ground truth 2D keypoints, 2) acceleration, and 3) differences to the original prediction. Table 2 and Fig. 5 show that it is possible to make the predicted trajectory smoother but is much harder to correct global motions *even with* GT 2D keypoints. In comparison, the motion prior learned by HaPTIC provides a better initialization for test-time optimization. PA-MPJPE stands for Procrustes-Aligned pose error.

4.2. Image-based Estimation

While HaPTIC is designed for 4D hand trajectories from videos, it generalizes to predict 3D hand poses from single images when video input is not available. Table 3 reports PCK at different thresholds on the HInt [46] benchmark. They are out-of-distribution datasets. The single-image hand predictions from HaPTIC outperform HaMeR and other baselines, whose margin is even more significant on occlusion subsplit (see *Sup. Mat.*). It is probably because the global CA layers are trained to capture full-frame context to infer hand poses, allowing plausible estimates even when the hand is occluded or blurry. We also report comparable 3D hand pose performance on other image-based benchmark [21] in *Sup. Mat.*.

4.3. Ablation Study

In Table 5, we analyze the effect of our design choices by comparing variants of our models that train and test on one allocentric view in ARCTIC-EXO. When we only train on video datasets without auxiliary image datasets, the model does not generalize well to the test split probably because the appearance of video training data is not diverse enough. Without cross-view attention (No Cross-View SA), the global trajectory performs the worst as it cannot fuse information from adjacent frames. Without attending to the global frame (No Global CA), the trajectory is less smoother (higher ACC-NORM) probably because the original frame, which varies less than hand crops in a sequence, stabilizes the predictions. We also try flipping the order of the two attention layers in the transformer decoder and find that the current design leads to the optimal global trajectory.

	GA-MPJPE	FA-MPJPE	ACC-NORM	PA-MPJPE
No Image Data	2.34	5.52	1.68	0.55
No Cross-View SA	2.82	6.63	**1.41**	0.65
No Global CA	2.25	5.12	1.62	0.54
Flip Order	2.20	5.23	1.62	**0.53**
Full Model	**1.83**	**4.27**	1.59	0.54

Table 5. **Ablation studies.** We compare our full model with variants that do not mix training batch with image data, do not have one of the attention layers, or have attention layers with flipped order on one view of ARCTIC-EXO.

Crop SubSplit	Method	@0.05	@0.1	@0.15
Center / Side	HaMeR	44.2 / 44.7	78.4 / 76.7	90.6 / 88.7
	HaPTIC	**47.7 / 45.9**	**80.3 / 78.7**	**91.8 / 89.9**
Flat / Square	HaMeR	46.9 / 42.1	79.2 / 75.9	90.6 / 88.8
	HaPTIC	**49.5 / 44.2**	**80.9 / 78.3**	**91.4 / 90.3**
Size L / M / S	HaMeR	50.8 / 45.8 / 35.7	82.3 / 79.1 / 69.8	92.3 / 90.9 / 84.7
	HaPTIC	**54.3 / 47.9 / 37.5**	**84.8 / 81.0 / 71.5**	**93.8 / 91.9 / 85.9**

Table 6. **Robust to Hand Crops**: HInt-VISOR test set is split by hand crop size, aspect ratio, and location.

Variation of Hand Crop. We break down the results of Table 3 in Table 6, based on the location, aspect ratio and size of the hand crops. HaPTIC outperforms HaMeR regardless of the crop variations. We also show in *Sup. Mat.* that HaPTIC is more robust to occluded joints.

4.4. In-the-Wild Results

Finally, we show more qualitative results in Fig. 6, including predictions from in-the-wild videos (top) and single images (bottom). HaPTIC can handle different hand sizes and different camera viewpoints. It can handle fast hand motions like drumming (row 5) while it correctly predicts the holding hand being static (row 2). The reconstruction also includes challenging cyclic motions like whisking (row 4). In the bottom row of Figure 1, the reconstructed hand trajectory desirably follows a straight line back and forth. The image reconstruction results show good 2D alignment of hand pose, even with severe hand truncation. We encourage readers to see *Sup. Mat.* videos.

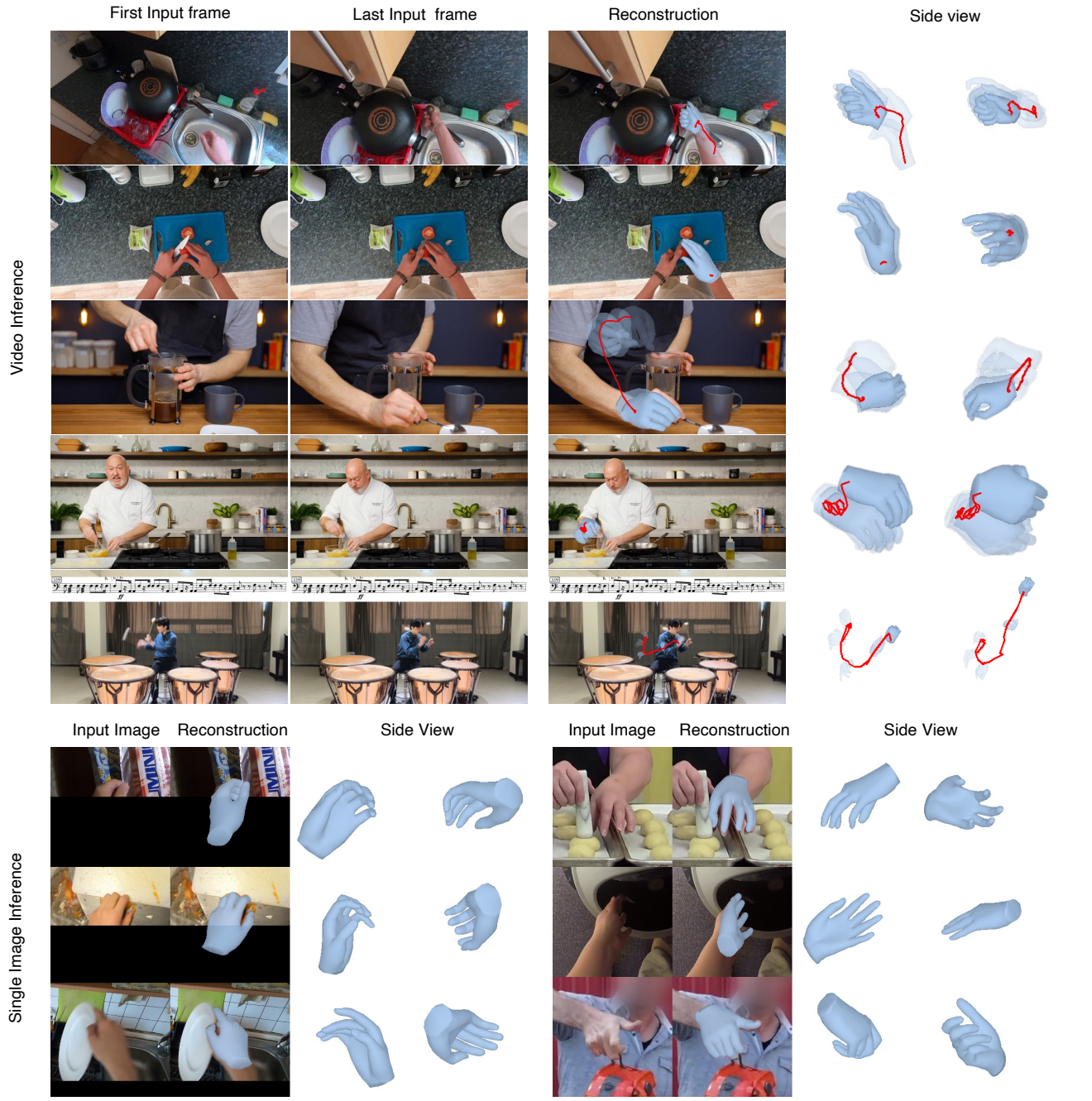

Figure 6. **Qualitative Results.** We show more qualitative results of our general model HaPTIC. Upper: Given videos from various datasets or from the Internet, we show reconstructed hand pose and trajectory from the original image view and from two side views. Lower: Given single images from the HInt dataset or the Internet, we show reconstructed poses from the original image view and side views.

5. Conclusion

In this work, we present HaPTIC, the first feed-forward method to predict 4D hand trajectory in global coordinates. To address the issue of limited video training data, we adapt a high-quality image-based model with novel attention layers and train it on both video and auxiliary image datasets. The adaption is simple and effective, including two types of attention and an additional prediction head. While we have achieved state-of-the-art results, the method assumes a reliable 2D hand tracking system. An interesting direction would be jointly tracking and reconstructing. Another possible direction is to infer hand trajectory in the world instead of camera coordinates. We hope that HaPTIC can serve as a useful tool for downstream tasks like hand-object interaction, robotics, and AR/VR.

References

[1] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. In *ICRA*, 2024. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 5

[3] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, 2023. 3

[4] A. Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *CVPR*, 2023. 3

[5] Manuel Caeiro-Rodríguez, Iván Otero-González, Fernando A Mikic-Fonte, and Martín Llamas-Nistal. A systematic review of commercial smart gloves: Current status and applications. *Sensors*, 2021. 2

[6] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 5

[7] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *NeurIPS*, 2022. 3

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 2

[9] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. HMP: Hand motion priors for pose and shape estimation from video. In *WACV*, 2024. 2

[10] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561*, 2023. 2

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 2, 5

[12] Hao-Shu 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. *TPAMI*, 2022. 5

[13] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *ICCV*, 2023. 2, 3

[14] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3D with multi-view diffusion models. *NeurIPS*, 2024. 3

[15] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*, 2018. 2

[16] Erik Gartner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3D human pose from monocular video. In *CVPR*, 2022. 3

[17] Erik Gartner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Pace: Physics-based animation and control of expressive 3D characters. In *CVPR*, 2022. 3

[18] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *ICCV*, 2023. 3

[19] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 3, 5

[20] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *COLM*, 2024. 2

[21] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 5, 7

[22] Yana Hasson, G"ul Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *3DV*, 2021. 2, 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *ICLR*, 2022. 3

[25] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. MVD-Fusion: Single-view 3D via depth-consistent multi-view generation. In *CVPR*, 2024. 3

[26] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 2, 5

[27] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 3

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 3

[29] Leyla Khaleghi, Alireza Sepas-Moghaddam, Josh Marshall, and Ali Etemad. Multiview video-based 3-d hand pose estimation. *TAI*, 2021. 2

[30] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 2, 5

[31] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *ECCV*, 2022. 3

[32] Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *CVPR*, 2024. 2

[33] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 3, 5

[34] Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Yebin Liu, Wei Jing, Qi Yan, Qianying Wang, and Hongwen Zhang. Omni-hands: Towards robust 4d hand mesh recovery via a versatile transformer. *arXiv preprint arXiv:2405.20330*, 2024. 2

[35] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 2, 7

[36] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2, 7

[37] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*, 2021. 2

[38] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022. 2

[39] Yuan Liu, Chu-Hsing Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *ICLR*, 2023. 3

[40] Jian Ma and Dima Damen. Hand-object interaction reasoning. In *AVSS*, 2022. 1

[41] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 2

[42] 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 *ECCV*, 2020. 2, 5

[43] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *CVPR*, 2022. 7

[44] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022. 2

[45] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 3

[46] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2, 3, 5, 7

[47] Siyou Pei, Alexander Chen, Jaewook Lee, and Yang Zhang. Hand interfaces: Using hands to imitate objects in ar/vr for expressive interactions. In *CHI*, 2022. 1

[48] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3D human motion model for robust pose estimation. In *PICCV*, 2021. 3

[49] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 3

[50] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. In *TOG*, 2017. 2, 3

[51] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3D whole-body pose estimation system via regression and integration. In *ICCVW*, 2021. 2, 3, 7

[52] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

[53] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and X. Yang. Mvdream: Multi-view diffusion for 3D generation. *ICLR*, 2023. 3

[54] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 2, 3, 5

[55] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 5

[56] Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2022. 3

[57] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Tao Mei, and Michael J Black. Monocular, one-stage, regression of multiple 3D people. In *PICCV*, 2021. 3

[58] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Trace: Monocular global human trajectory estimation in the wild. In *CVPR*, 2023. 3

[59] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3D hand-mesh reconstruction. In *ICCV*, 2021. 2

[60] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic Fields: Marrying 3D geometry and video understanding. *NeurIPS*, 2024. 2

[61] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *RSS*, 2024. 2

[62] Ruocheng Wang, Pei Xu, Haochen Shi, Elizabeth Schumann, and C Karen Liu. F\" urelise: Capturing and physically synthesizing hand motions of piano performance. *SIGGRAPH Asia*, 2024. 2

[63] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. *TOG*, 2009. 2

[64] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 2, 5

[65] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3D hand pose and shape estimation. In *ECCV*, 2020. 2

[66] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 5, 6

[67] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 3

[68] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, 2024. 2

[69] Brent Yi, Vickie Ye, Maya Zheng, Lea M"uller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 2

[70] Paper Authors Your. Slahmr: Simultaneous localization and human mesh recovery. In *CVPR*, 2023. 5

[71] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *TOG*, 2021. 3

[72] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dynhamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27716–27726, 2025. 2

[73] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 3, 5

[74] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. *arXiv preprint arXiv:2501.02973*, 2025. 2, 6

[75] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, 2019. 2

[76] Zhishan Zhou, Zhi Lv, Minqiang Zou, Yao Tang, Jiajun Liang, et al. A simple baseline for efficient hand mesh reconstruction. In *CVPR*, 2024. 2

[77] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single rgb images. In *ICCV*, 2017. 5

[78] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *PICCV*, 2019. 2, 5