

Towards Egocentric 3D Hand Pose Estimation in Unseen Domains

Wiktor Mucha^{1,2} Michael Wray³ Martin Kampel¹

¹Computer Vision Lab, TU Wien ²SoftServe Inc., Kraków, Poland, ³University of Bristol

{wiktor.mucha, martin.kampel}@tuwien.ac.at, michael.wray@bristol.ac.uk

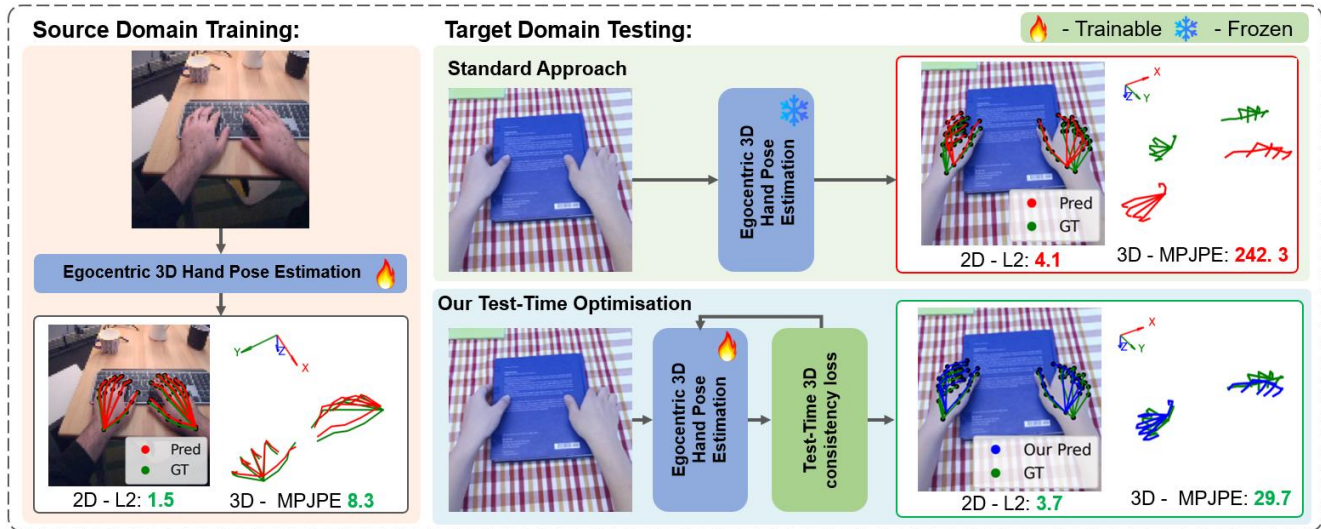


Figure 1. **V-HPOT improves egocentric 3D hand pose estimation in cross-domain scenarios**, where a model is trained in source and tested in different target domain. We propose training the model to estimate the z -coordinate in virtual camera space to mitigate overfitting to source domain camera intrinsics. At test time, we optimise predictions using a novel self-supervised loss that enforces 3D spatial consistency, allowing the model to refine depth perception by aligning depth-augmented poses in 3D camera space. This significantly improves absolute 3D pose error (MPJPE) without any labelled data, outperforming existing methods focused solely on 2D improvement.

Abstract

We present *V-HPOT*, a novel approach for improving the cross-domain performance of 3D hand pose estimation from egocentric images across diverse, unseen domains. State-of-the-art methods demonstrate strong performance when trained and tested within the same domain. However, they struggle to generalise to new environments due to limited training data and depth perception – overfitting to specific camera intrinsics. Our method addresses this by estimating keypoint z -coordinates in a virtual camera space, normalised by focal length and image size, enabling camera-agnostic depth prediction. We further leverage this invariance to camera intrinsics to propose a self-supervised test-time optimisation strategy that refines the model’s depth perception during inference. This is achieved by applying a 3D consistency loss between predicted and in-space scale-transformed hand poses, allowing the model to adapt to target domain characteristics

without requiring ground truth annotations. *V-HPOT* significantly improves 3D hand pose estimation performance in cross-domain scenarios, achieving a 71% reduction in mean pose error on the H2O dataset and a 41% reduction on the AssemblyHands dataset. Compared to state-of-the-art methods, *V-HPOT* outperforms all single-stage approaches across all datasets and competes closely with two-stage methods, despite needing $\approx \times 3.5$ to $\times 14$ less data. <https://github.com/wiktormucha/vhpot>.

1. Introduction

Egocentric vision has emerged as a rapidly growing research area in computer vision, fueled by datasets including *Epic-Kitchens* [7] *Ego4D* [12] or *HOT3D* [1], with applications spanning assistive technologies [26] and robotics [11]. Since hands play a central role in the egocentric viewpoint, one of the key challenges is to accurately estimate the 3D hand pose of the camera wearer. A correct

3D hand pose representation enables interaction with virtual environments in VR/AR applications, facilitates action recognition, or can be used to transfer manipulation skills from human experts to robots in real-world scenarios [1].

Recent works tackle egocentric 3D hand pose estimation through two distinct approaches. Two-stage methods [23, 28, 33] first detect hand regions then estimate keypoints, allowing detector pre-training on large datasets without 3D labels. However, this introduces critical limitations: increased computational overhead, cascading failures when detection errors occur, and loss of crucial contextual information in hand-object interactions due to tight cropping [6]. Single-stage methods [6, 27, 40] overcome these limitations by directly estimating keypoints from full RGB frames in one forward pass, preserving scene context and enabling joint reasoning about hands and objects [13, 17, 43], which is essential for accurately capturing the complex spatial relationships in interaction scenarios.

Despite these advancements, most methods only report performance within the same domain [6, 18, 27, 28, 40], leaving performance in unseen domains unaddressed – critical for real-world applications that must function robustly in uncontrolled environments. Models overfit to specific camera intrinsics and environments, causing performance degradation in novel settings. One possible solution is to train models with larger and more diverse datasets; however, acquiring 3D pose annotations is costly, requiring multi-camera setups [18, 28] or motion capture systems [1] that constrain data collection to laboratories, limiting generalisation capabilities towards real-world data.

We address these challenges with V-HPOT (Virtual Camera-based Hand Pose Optimisation at Test-time) presented in Figure 1, a single-stage approach with two key components. First, we formulate z-coordinate prediction in Virtual Camera (VC) [2] space rather than metric space, making depth perception less dependent on camera intrinsics and enabling depth augmentations that enhance generalisation across datasets with varying focal lengths. During training, we incorporate an auxiliary pseudo-depth prediction task supervised by *DPT-Hybrid* [34], improving cross-domain generalisation. Second, we introduce a novel self-supervised Test-Time Optimisation (TTO) strategy that dynamically adapts the model to previously unseen environments by refining the feature extractor through a consistency loss between predicted and depth-transformed hand poses in 3D space. This is only possible with our camera-agnostic depth representation, which maintains consistency under depth transformations, unlike conventional metric representations that suffer from scale-depth ambiguity.

V-HPOT achieves consistent improvements in cross-domain scenarios, reducing mean pose error by 71% and 41% for *H2O* [18] and *AssemblyHands* [28] datasets, respectively. Compared to the state of the art, we outper-

form other single-stage methods in cross-domain settings on *H2O* [18] and *AssemblyHands* [28] across all evaluation metrics. Additionally, we assess the 2D projections of our method on the in-the-wild *Epic-Kpts* dataset [33], which lacks 3D annotations, where it surpasses all single-stage approaches and performs comparably with two-stage methods, despite our model being trained from $\approx \times 3.5$ to $\times 14$ less data. Our key contributions are as follows:

- We adapt virtual camera space to hand pose estimation, reducing depth perception’s dependency on camera intrinsics and enabling effective depth augmentations.
- We propose pseudo-depth prediction as an auxiliary task for cross-domain 3D hand pose estimation.
- We propose a test-time optimisation framework with a novel self-supervised 3D consistency loss for 3D hand pose estimation, leveraging an intrinsic invariant virtual camera space. It allows the model to refine depth perception in unseen domains without any labelled data by enforcing alignment between 3D depth-augmented poses.
- We demonstrate V-HPOT effectiveness through extensive experiments across multiple dataset domains, achieving substantial cross-domain performance improvements: 71% error reduction on the RGB *H2O Dataset* and 41% on the monochromatic *AssemblyHands*.

2. Related Work

Egocentric 3D Hand Pose Estimation. Hand pose estimation in egocentric vision is challenging due to self-occlusion, limited field of view, and varying perspectives. Current approaches can be categorised into two main paradigms: Two-stage methods [30, 31, 33, 35] first localise hand regions before estimating poses. FrankMocap [35] employs a convolutional backbone and an HMR-style [16] decoder to predict MANO parameters under 2D and 3D supervision. HandOccNet [30] integrates an FPN [22] backbone and transformer modules for heatmap intermediate representations. HaMeR [31] leverages a ViT [9] with a transformer decoder. WildHands [33] incorporates image features and intrinsics-aware positional encodings for each hand crop, before regressing 3D hand pose. Auxiliary supervision comes from hand segmentation masks and grasp labels in the *Epic-Kitchens* dataset. While effective, these approaches introduce computational overhead and lose crucial interaction context due to cropping [6] and train on a large amount of data.

In contrast, single-stage methods estimate hand pose directly from entire images [6, 10, 18, 27, 39, 40]. Tekin et al. [39] predict a 3D pose grid from a single RGB image, while Kwon et al. [18] extend this to both hands. Transformer-based models have also been explored, with Cho et al. [6] focusing on frame-wise 3D reconstruction and Wen et al. [40] addressing occlusions with a sequence-based approach. ArcticNet-SF [10] combines a convolu-

tional backbone with an HMR-style decoder for hand and object pose estimation. SHARP [27] introduces pseudo-depth estimation to isolate hands and objects in egocentric scenes, leveraging arm distance for improved accuracy. Many of these methods extend to 3D object pose estimation, where joint reasoning about hands and objects is proven to enhance reconstruction of both [13, 17, 43].

Despite advancements, most studies focus on controlled settings [6, 18, 28, 39, 40], with limited attention to real-world egocentric data [31, 33]. Building on single-stage approaches [6, 18, 40], we evaluate 3D hand pose in a cross-domain setting with different sensors following [33]. Compared to [33], our method improves the performance of single-stage approaches, relies on a single training dataset, minimises the number of auxiliary tasks to one introducing a novel one, and directly improves depth estimation (z -coordinate) from a single RGB image.

Test-time optimisation for pose estimation. Test-time optimisation improves model performance during inference by adapting to test data in an unsupervised manner [15, 32, 37]. This adaptation occurs online, either by optimising an auxiliary task [37] or passively through non-training methods such as statistical adaptation of Batch Normalisation layers [19]. Test-time adaptation for pose estimation is achieved via self-training and self-supervision tasks. Li et al. [20] use multi-view images of the same person to synthesise the input image based on initial pose prediction. Hu et al. [15] employ initial pose predictions to mask image regions and apply inpainting as an auxiliary task. Hasson et al. [14] perform weakly supervised learning with differentiable rendering and optical flow for temporal consistency in 3D hand-object reconstruction, while [3] refine hand-object alignment using additional knowledge regarding 3D contact priors and collision avoidance.

Our work differs from existing methods as we focus on 3D hand pose estimation, rather than simply refining 2D predictions [15, 20, 32]. Current 3D hand pose adaptation approaches either rely on additional ground truth data [14] or require object interaction cues [3], which limits their applicability in non-interaction scenarios. In contrast, we propose a novel self-supervision for enhancing depth perception, leading to more robust and generalisable 3D hand pose estimation across domains, independent of camera intrinsic, without the need for additional ground truth labels.

Self-supervision in hand pose. TTO heavily depends on selecting an unsupervised task aligned with the main objective. Various self-supervised approaches exist for hand pose estimation. Some generate pseudo-labels using 2D pose estimators [5, 44] or utilise 2D ground truth poses [4], refining them to improve 3D representations through cross-sample consistency. HaMuCo [44] leverages multi-view consistency during training to enhance single-view estimates. TASSN [4] enforces temporal consistency be-

tween forward and reverse video frames, ensuring stable 3D predictions from single-view ground truth 2D labels. For single-view, non-temporal approaches, S2-Hand [5] refines noisy 2D labels through rendering-based self-supervision, while Spurr et al. [36] apply contrastive learning with geometric and appearance-based augmentations between similar frames, enforcing consistency in the 2D space, while Yang et al. [41] utilise rotations. Lin et al. use contrastive pretraining by mining similar hand representations [21].

Our work introduces a novel self-supervised task specifically for TTO, unlike prior approaches focused on pre-training. We neither rely on pseudo-labels [5, 21, 44], temporal cues [36], nor rendering [5]. While we leverage consistency between similar poses [4, 36], we avoid multi-frame dependencies [4] or data mining [21] not suitable for test-time. Additionally, we extend beyond image-based augmentations [36] by applying transformations directly in 3D space to augment hand poses and force 3D consistency.

3. Method

Our method introduces two key components to improve the egocentric 3D hand pose estimation (see Figure 2). (1) We train the model to estimate depth (z -coordinate) in virtual camera space, making predictions agnostic to camera intrinsic variations across datasets while introducing pseudo-depth estimation as an auxiliary task. (2) We propose a test-time training pipeline that enhances the model’s adaptability during inference by incorporating a 3D consistency loss that leverages various distance augmentations in the virtual camera space, enabling the model to learn robust hand pose estimation in new domains in a fully unsupervised manner.

3.1. 3D Hand Pose Estimation Network

Our egocentric hand pose estimation network adopts a single-stage architecture. Following standard practice, we represent each hand with 21 keypoints $j \in J$ denoting joint positions, expressed as $P_{L,R}^{3D}(x, y, z) \in \mathbb{R}^{J \times 3}$ where L, R stands for the left or right hand and x, y, z represents the point j in 3D. We follow a common practice, where x, y are estimated via heatmaps and z is regressed [27].

Given an input RGB frame $I_n^{RGB} \in \mathbb{R}^{3 \times 224 \times 224}$, we extract features using *EfficientNetV2-S* [38] to produce a feature map F_M . This backbone is selected following experiments in [25], where it outperforms other architectures in the egocentric hand pose task. F_M is then fed into two independent upsamplers—one per hand—to generate 2D keypoint heatmaps $H_{L,R}$. Per-hand upsamplers are first presented in InterHands [24], a pioneering study focused on interacting hands, and became common in egocentric hand pose [6, 18, 27, 40] due to improvements in pose estimation during hand interactions. Each upsampler consists of four transposed convolution layers with batch normalisation and ReLU, followed by a pointwise convolution in the final

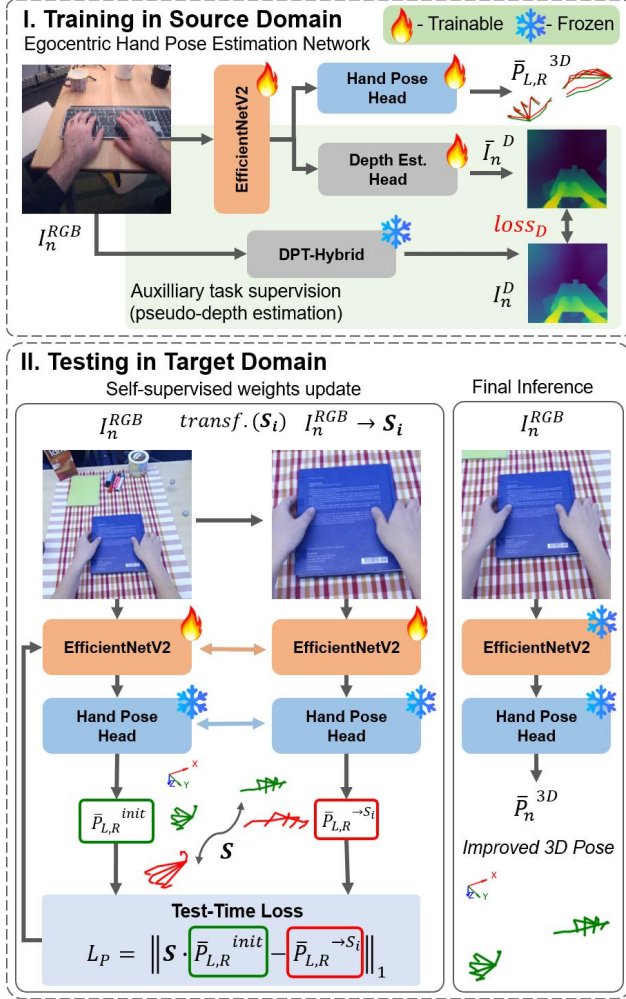


Figure 2. **V-HPOT**: We train in source domain to estimate 3D hand pose $\bar{P}_{L,R}^{3D}$ from an image I_n^{RGB} in virtual camera space, with an auxiliary pseudo-depth task I_n^D supervised by *DPT-Hybrid* [34]. At test time, the model predicts an initial pose $\bar{P}_{L,R}^{init}$ and applies depth augmentation by scaling I_n^{RGB} with S_i in virtual camera space, yielding $\bar{P}_{L,R}^{S_i}$. A 3D consistency loss between these poses refines the backbone, enhancing final predictions.

layer. The output, $H_{L,R} \in \mathbb{R}^{J \times 112 \times 112}$, represents per-joint probability distributions, which are converted into 2D keypoints ($P_{L,R}^{2D} \in \mathbb{R}^{J \times 2}$). Depth (z) is predicted separately via a Multi-Layer Perceptron (MLP) ($MLP_{L,R}^Z$), and combined with $P_{L,R}^{2D}$ to form 2.5D keypoints ($P_{L,R}^{2.5D}$). To obtain the 3D hand pose in camera space, we apply the pinhole camera model (Equation 1), transforming $P_{L,R}^{2.5D}(x, y, z)$ into $P_{L,R}^{3D}$ using intrinsic parameters K .

$$P_{L,R,j}^{3D} = z_j \cdot K^{-1} \begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix}, \quad \forall j \in J \quad (1)$$

Additionally, hand presence is determined via handedness modules ($MLP_{L,R}^H$), predicting the probability of each

hand being present ($h_L, h_R \in \mathbb{R}^2$) in the image I_n^{RGB} .

3.2. Understanding depth in new domains

Introducing virtual camera space. Estimating metric depth from pixel values requires an implicit mapping of 2D keypoints to 3D distances, which becomes ambiguous due to variations in camera intrinsics. Since camera intrinsics differ across datasets, models trained on a single dataset tend to overfit to specific depth representations, leading to poor generalisation in cross-domain scenarios where data is often captured by different sensors.

To address this issue, we implement a concept of a *virtual camera (VC)* [2] for hand pose estimation. This approach transforms metric depth values z to a uniform virtual camera space, denoted as z_v . The transformation is achieved by scaling depth values z using known camera intrinsics, ensuring a consistent, effective image size and focal length across datasets. We denote these standardised parameters as H_v and f_v , which are treated as hyperparameters. Following this, virtual depth is defined as the transformation of ground truth z -coordinate $z_v = z \cdot \frac{f_v}{f} \frac{H}{H_v}$ where H and f are source image height and focal length. The model learns to predict $\hat{z}_v$, which is scaled back to metric z for the output. Additionally, leveraging virtual camera space enables scale augmentations that enhance both 2D and 3D pose estimation. Without it, augmentations would improve 2D representation but degrade 3D understanding [2].

Training objectives. Our objectives are 3D hand pose estimation and pseudo-depth estimation as an auxiliary task. The prediction of a pseudo-depth map is supervised using the prediction by the *DPT-Hybrid* [34]. We use an L1 loss over all pixels N to minimise the difference between the predicted depth $\hat{I}^D$ and the reference pseudo-depth I^D :

$$\mathcal{L}_d = \frac{1}{N} \sum_{i=1}^N |\hat{I}_i^D - I_i^D| \quad (2)$$

For 2D hand keypoint localisation, we supervise the network using an Intersection over Union (IoU) loss over the predicted heatmaps. Given the predicted heatmap $\hat{H}$ and the ground truth heatmap H , the IoU loss is defined as:

$$\mathcal{L}_{xy} = 1 - \frac{\sum_{j=1}^{21} \min(\hat{H}_j, H_j)}{\sum_{j=1}^{21} \max(\hat{H}_j, H_j)} \quad (3)$$

For 3D hand pose estimation, we directly optimise the obtained depth (z_v -coordinate) using an L1 loss:

$$\mathcal{L}_{z_v} = \sum_{j=1}^{21} |\hat{z}_{v,j} - z_{v,j}| \quad (4)$$

To predict whether the image contains left, right, or both hands, we use a handedness classification module. The

handedness prediction $\hat{h}$ is supervised using the cross-entropy loss with the ground truth label h :

$$\mathcal{L}_h = - \sum_{c \in \{\text{left}, \text{right}\}} h_c \log \hat{h}_c \quad (5)$$

where h is a one-hot encoded ground truth vector indicating whether the hand is left or right. The final objective is:

$$\mathcal{L}_{train} = \lambda_d \mathcal{L}_d + \lambda_{xy} \mathcal{L}_{xy} + \lambda_{z_v} \mathcal{L}_{z_v} + \lambda_h \mathcal{L}_h \quad (6)$$

where $\lambda_d, \lambda_{xy}, \lambda_{z_v}, \lambda_h$ are hyperparameter weights controlling the relative importance of each term.

Generalising to new domains at test time. We propose a fully self-supervised test-time optimisation pipeline that adapts hand pose estimation to unseen environments without requiring target domain annotations. This adaptation during inference is only possible because of our virtual camera transformation, which creates a camera-intrinsic-uniform representation space.

Specifically, we introduce a consistency loss to improve 2D and depth generalisation. The model first infers the 3D hand pose in virtual camera space, producing an initial prediction $\mathbf{P}_{L,R}^{\text{init}}$. We then apply $n = 2$ independent random scaling factors $S_i \sim \mathcal{U}(1.0, 1.25)$ to augment the depth, generating n transformed poses $\mathbf{P}_{L,R}^{\rightarrow S_i}$ that simulate shifts in camera distance. These transformed poses are subsequently passed through the network again, producing new predictions. Since the spatial relationships between joints should remain consistent despite the depth transformation, we enforce this 3D consistency via the L1 loss function:

$$\mathcal{L}_{\text{test-time}} = \sum_{i=1}^n \left\| S_i \cdot \mathbf{P}_{L,R}^{\text{init}} - \mathbf{P}_{L,R}^{\rightarrow S_i} \right\|_1 \quad (7)$$

This consistency loss optimises the feature extractor weights while keeping the hand pose head weights frozen. It is carried out *online* using the first F_N frames of the test data. This choice ensures a balance between quick adaptation and avoiding overfitting, allowing the model to adjust to domain-specific variations without compromising generalisation. Empirical observations show that using a larger fraction leads to diminishing returns, while a smaller fraction results in insufficient adaptation. Initially, the model is trained on the source dataset and undergoes a one-time weight initialisation. During testing, the weights are iteratively refined using $\mathcal{L}_{\text{test-time}}$ before making predictions in the same manner for every domain. After processing the first F_N of the data, optimisation is disabled, and the model proceeds with inference without any further updates.

4. Experiments

We conduct cross-domain evaluation: training on source dataset, testing on target dataset from different domain.

4.1. Experimental Methodology

Datasets. Our study employs four egocentric datasets: three from laboratory settings and one collected in the wild. For lab datasets, we use: (1) *H2O* [18], which provides multi-view RGB-D recordings with 21 keypoint annotations per hand, 6D object poses, and scene metadata (55K training, 23K testing frames); (2) *AssemblyHands* [28], a large-scale monochrome dataset (380K training, 32K validation frames); and (3) *HOT3D* [1], featuring multi-view RGB video streams of 19 subjects interacting with 33 objects (170K training, 42K testing frames). For in-the-wild evaluation, we use *Epic-Kpts* [33], a subset of Epic-Kitchens [7] with 2D annotations for 21 hand joints from 4K images sampled from VISOR split [8]. For this dataset, we project our 3D predictions into 2D for assessment. Specific dataset combinations for experiments are detailed in later sections.

Metrics. We use standard hand pose estimation metrics: Mean Per Joint Position Error (MPJPE), measuring Euclidean distance (mm) between predicted and ground-truth 3D keypoints across 21 hand joints; root-aligned MPJPE (MPJPE-RA) (mm) with poses aligned by root joint subtraction; Mean Relative-Root Position Error (MRRPE) (mm), the quantifying distance between the left and right hand root joints and measuring absolute pose information; and L2 error (pixels) between ground truth 2D keypoints and 2D projections of predicted 3D keypoints on 224×224 input images. All 3D metrics are computed in camera space.

Experimental Setup. Each run, including ablation studies, is repeated three times to mitigate the impact of random network initialisation. To enable comparison with related work [6, 31, 33, 35], the model is trained on *HOT3D* [1] as source dataset and evaluated on *H2O Dataset* [18] and *AssemblyHands* [28]. A cross-domain experiment showing improvements in all scenarios is available in the Supp. A.2. The optimisation is performed using Stochastic Gradient Descent (SGD) with a learning rate of $l_r = 0.1$ and momentum $m = 0.9$. The learning rate is progressively reduced by a factor of $\alpha = 0.5$ every five epochs, starting from the 25th epoch, where the process ends at 50 epochs. Training loss $\mathcal{L}_{train}$ parameters are set up following [27]. For virtual camera parameters, we select constant $H_v = 720$ and $f_v = 512$ following [2]. Data augmentation includes horizontal flipping, resizing, and appearance-based transformations such as blurring and de-pixelization. The batch size is set to $b_s = 32$. Model weights are saved based on the lowest MPJPE observed in the validation set.

During test, we optimise $\mathcal{L}_{\text{test-time}}$ using SGD with $l_r = 0.3$ and $m = 0.2$. For experiments, we use F_N equal 5% of test data, whilst for comparison with other methods we use a fixed number (5% of *H2O* - 960 frames) to simulate realistic deployment scenarios. The model performs best when enforcing consistency loss across $n = 2$ shifted poses.

4.2. Results

We present ablations on V-HPOT’s key components: VC, TTO, auxiliary task, and TTO loss. Finally, we compare V-HPOT with the state of the art in cross-domain scenarios.

Virtual Camera and Test-Time Optimisation. We evaluated our VC approach and TTO impact through ablations:

1. *base*: Training and testing without VC transformation
2. *VC*: Applying VC transformation during training for camera intrinsic robustness, without TTO
3. *TTO*: Incorporating TTO without VC
4. *VC+TTO*: Integrating both VC transformation and TTO, enabling intrinsic-independent depth learning with adaptive inference (our complete V-HPOT architecture)

Results in Table 1 demonstrate that V-HPOT (*VC+TTO*) yields superior performance across all metrics, reducing MPJPE by 70% on H2O and 41% on AssemblyHands (AsHa). The combined VC and TTO enhances depth perception in unseen domains by addressing key limitations: *TTO* alone leads to source camera intrinsic overfitting, while *VC* alone lacks a feedback mechanism for weight adjustment. Both target datasets exhibit improvements with our best (*VC+TTO*) in 3D error, but gains vary by dataset. *AssemblyHands* [28] showed greater 2D L2 reduction, likely due to its monochromatic nature offering more improvement potential compared to the source RGB (HOT3D). In *Epic-Kpts* L2 reduces by 30% (12.9-8.4). Figure 3 visualises coordinate-specific losses (L2 for x, y ; L1 for z) on target datasets, comparing *base* model with (*V-HPOT*). Results show consistent improvements across images, even with high error in initial prediction.

Auxiliary pseudo-depth estimation. We evaluate its impact on training in Table 2 by comparing the *V-HPOT* against *V-HPOT- D_{tr}* , a version with excluded auxiliary pseudo-depth estimation task (D_{tr}) during training and with *V-HPOT(DA)* where we utilise DepthAnything [42] instead of *DPT-Hybrid*. Results show that including D_{tr} improves cross-domain performance across all metrics and datasets. Performance between *V-HPOT(DA)* and *V-HPOT* varies, thus we select *DPT-Hybrid* as it achieves superior absolute 3D error and faster inference (26.8ms vs 39.8ms mean of 1000 trials on RTX3090). Qualitative observations show that DepthAnything estimates finer background details, whilst hands remain similar (Supp. A.8).

Test-Time Optimisation Loss. We evaluated our TTO loss $\mathcal{L}_{\text{test-time}}$ through ablation studies:

1. *base*: No test-time optimisation
2. $\mathcal{L}_{xy}$: Consistency loss between $\mathbf{P}_{L,R}^{\text{init}}$ and augmented pose $\mathbf{P}_{L,R}^{\rightarrow S_1}$ on 2D coordinates only, enforcing keypoint alignment in image space without considering depth.
3. $\mathcal{L}_z$: Consistency loss on z only between $\mathbf{P}_{L,R}^{\text{init}}$, $\mathbf{P}_{L,R}^{\rightarrow S_1}$
4. $\mathcal{L}_{xyz}$: Combined consistency across all x, y, z coordinates between $\mathbf{P}_{L,R}^{\text{init}}$, $\mathbf{P}_{L,R}^{\rightarrow S_1}$ poses

Table 1. **Ablation study presenting impact** of proposed virtual camera transformation and TTO strategy. We observe an improvement of V-HPOT (*VC+TTO*) over baseline in every metric.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
<i>base</i>	179.6	297.7	52.7	105.3	126.2	301.8	7.4	80.1
<i>VC</i>	146.1	302.3	63.9	105.8	76.9	288.2	7.2	84.0
<i>TTO</i>	209.6	261.8	59.7	105.9	139.6	343.0	6.5	64.2
<i>VC+TTO</i>	53.3	174.5	51.1	90.6	54.1	253.2	5.8	36.9

Table 2. **Impact of auxiliary pseudo-depth estimation D_{tr} .**

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
<i>V-HPOT - D_{tr}</i>	60.6	237.7	56.1	98.1	78.9	310.1	5.9	56.1
<i>V-HPOT(DA)</i>	67.2	200.9	44.1	86.1	77.8	158.6	5.6	43.9
<i>V-HPOT</i>	53.3	174.5	51.1	90.6	54.1	253.2	5.8	36.9

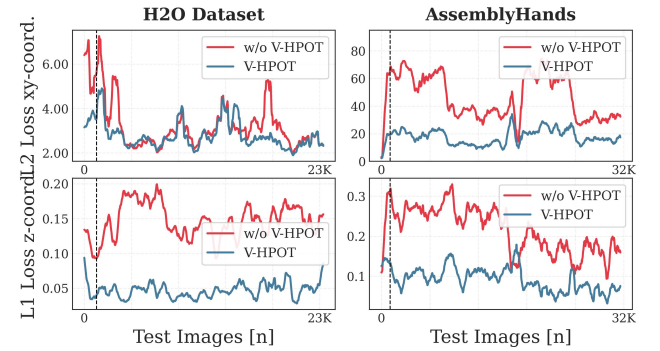


Figure 3. **Visualisation of L2 loss for x, y and L1 loss for the z coordinate over test samples in cross-domain scenarios.** The dashed line represents 5% of data where V-HPOT stops optimising. Continuous improvement is observed for most samples.

5. $\mathcal{L}_{xyz}^{n=2}$: We extend the consistency loss from one augmented pose ($n = 1$) to two augmented poses ($n = 2$) ($\mathbf{P}_{L,R}^{\text{init}}, \mathbf{P}_{L,R}^{\rightarrow S_1}, \mathbf{P}_{L,R}^{\rightarrow S_2}$)
6. $\mathcal{L}_{xyz}^{n=2} + D_t$: Adding pseudo-depth estimation to 5

Findings in Table 3. indicate that incorporating $\mathcal{L}_{xy}$ improves 2D estimations in both target datasets. However, while the 2D error decreases, MPJPE increases in *H2O* [18] and decreases in *AssemblyHands* [28], suggesting that 2D improvements alone do not guarantee better 3D estimation. Applying $\mathcal{L}_z$ results in lower MPJPE and 2D L2 error across both datasets. Combining both losses in $\mathcal{L}_{xyz}$ further improves MPJPE and L2 error in *H2O* [18] while also outperforming the *base* model in *AssemblyHands* [28].

The $\mathcal{L}_{xyz}^{n=2}$ variant achieves the best results on both datasets, demonstrating the effectiveness of our virtual camera approach with multi-view consistency enforcement.

Table 3. **TTO strategies impact.** Proposed depth-augmented 3D pose consistency loss achieves best results.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
<i>base</i>	179.6	297.7	52.7	105.3	126.2	301.8	7.4	80.1
$\mathcal{L}_{xy}$	226	177.5	79.4	93.3	95.88	271.2	6.2	40.1
$\mathcal{L}_z$	54.7	253.2	49.1	124.5	62.5	338.7	6.1	64.7
$\mathcal{L}_{xyz}$	53.99	190.9	50.9	94.4	57.4	253.9	5.9	43.6
$\mathcal{L}_{xyz}^{n=2}$	53.3	174.5	51.1	90.6	54.1	253.2	5.8	36.9
$\mathcal{L}_{xyz}^{n=2}+D_t$	56.5	175.3	52.7	91.5	54.2	254.7	5.9	36.7

Notably, incorporating pseudo-depth estimation at test time ($\mathcal{L}_{xyz}^{n=2}+D_t$) slightly increases the pose error, leading us to exclude it from the TTO process and use it only for training.

Comparison with other TTO approaches. As the fundamental difference between TTO methods for hand pose estimation lies in the self-supervision task, we create a unified TTO process and compare V-HPOT to [36], introducing 2D consistency loss, and [15] with image inpainting to *base* with excluded D_{tr} . Our comparison excludes methods [3, 14] which require object information or ground truth, or multi-view input [20]. V-HPOT uniquely targets 3D understanding, achieving reductions in both 2D and 3D errors, whilst competing methods only show 2D gains (Table 4). While [36] also relies on consistency loss, it is limited to 2D keypoint transformations. V-HPOT’s novelty lies in enforcing 3D spatial consistency, making it the first method to effectively handle single-camera 3D hand pose estimation with superior performance over [15, 36].

Comparison with foundation model DINOv2. In Table 5 we compare DINOv2 [29] cross-domain generalising capabilities with frozen weights (*D2*), fine-tuned last 4 layers (*D2-FT*), and as V-HPOT backbone (*V-HPOT-D2*). Results show *D2* and *D2-FT* cross-domain generalisation is insufficient in our task. As backbone, performance is comparable but inferior to *EfficientNetv2s* while being larger (86M vs. 21M parameters) and inferring $\approx 5\times$ slower (109.1 vs. 22.8ms mean of 1000 trials per batch on RTX3090).

4.3. Comparison with the State of the Art

We compare our approach to other egocentric 3D hand pose estimation methods in a cross-domain setting, using MPJPE-RA for 3D pose accuracy and MRRPE for absolute 3D error, following [33]. Our method outperforms other single-stage approaches, including H2OTR [6] and ArcticNet-SF [33], in every scenario, including in-the-wild Epic-Kpts [33], despite being trained on less data (Table 6).

Direct comparisons between V-HPOT and two-stage methods disadvantage our approach for three critical reasons: (1) two-stage methods utilise hand detectors trained

Table 4. **Comparison of different TTO/self-supervised strategies.**

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
<i>base</i>	139.1	329.2	53.2	112.6	111.1	463.3	8.0	88.3
<i>PeCLR</i> [36]	106.1	384.3	56.0	111.5	95.9	448.1	7.4	57.8
<i>MeTTA</i> [15]	142.2	349.9	53.3	108.5	112.3	413.9	7.8	86.1
V-HPOT	53.3	174.5	51.1	90.6	54.1	253.2	5.8	36.9

Table 5. **Foundation model DINOv2** as a backbone in frozen and fine-tuned setup, and with our V-HPOT TTO pipeline.

	MPJPE ↓		MPJPE-RA ↓		MRRPE ↓		L2 ↓	
	H2O	AsHa	H2O	AsHa	H2O	AsHa	H2O	AsHa
<i>D2</i>	226.7	259.7	78.8	151.1	160.7	468.6	32.8	48.7
<i>D2-FT</i>	185.7	249.8	58.1	132.1	127.8	426.2	7.9	45.8
<i>V-HPOT-D2</i>	59.4	191.2	58.0	93.2	64.1	221.8	7.5	44.3
V-HPOT	53.3	174.5	51.1	90.6	54.1	253.2	5.8	36.9

on massive, diverse datasets while we rely solely on limited egocentric data; (2) they process high-resolution hand crops, whereas we handle entire scenes at 224×224 pixels with hands occupying only a minimal area; and (3) they leverage larger training datasets. Despite these disadvantages, V-HPOT outperforms FrankMocap across all datasets while being trained on $\approx 4\times$ less data. We surpass HaMer on two datasets in MRRPE (absolute 3D error) with $\approx 14\times$ less training data and trail WildHands by only 5mm on H2O, which uses $\approx 3.5\times$ more training data. Most notably, our 2D L2 error is merely 1.2 pixels behind WildHands in *Epic-Kpts*, despite our whole-scene approach versus their focused hand crops, showing the strength of VC and our TTO in improving cross-domain hand pose performance.

4.4. Qualitative Results

We present qualitative comparisons with and without V-HPOT in Figure 4, showing 2D pose projections and corresponding 3D poses. Images (a–d) demonstrate V-HPOT improvements, while (e) shows a negative impact of V-HPOT. In *H2O Dataset* [18], quantitative errors are lower than in *AssemblyHands* [28], which is also reflected in the qualitative results. In *H2O*, V-HPOT performs well in scenarios where the subject interacts with objects, with the most significant improvements observed in wrist positioning. We identify failure cases in both datasets when hands move close to or directly towards the camera (Figure 4–e). Such close-camera scenarios are outliers in typical egocentric datasets, which focus on object manipulation. Therefore, improvements to data collection are required. Additionally, *AssemblyHands* contains stronger distortions, where V-HPOT tends to fail more frequently (Figure 4–e), which could be addressed with dedicated distortion-aware augmentations or normalising field of view across datasets.

Table 6. **Comparison of state-of-the-art egocentric hand pose models in cross-domain settings.** The table indicates whether a model follows a single-stage approach, the number of training domains and images, where the number in brackets for [10, 33] relates to the model tested on *AssemblyHands*. MPJPE-RA and MRRPE are reported in millimetres. * means results are referenced from [33].

Method	Single Stage	T. Domains	T. Img	H2O		Assembly		Epic-Kps
				MPJPE-RA	MRRPE	MPJPE-RA	MRRPE	L2
FrankMocap [35]	✗	6	675K	58.51	-	97.59	-	13.33
HaMeR [31]	✗	11	2.7M	23.82	147.87	45.49	334.52	4.56
WildHands [33]	✗	3	627K(267K)	31.08	49.49	80.40	148.12	7.20
ArcticNet-SF [10]*	✓	3	627K(267K)	83.84	325.55	110.76	326.94	35.02
H2OTR [6]	✓	1	55K	115.16	418.59	137.25	543.28	14.46
SHARP [27]	✓	1	180K	52.40	98.08	103.84	282.21	14.41
V-HPOT (Ours)	✓	1	180K	51.11	54.07	92.09	252.85	8.45

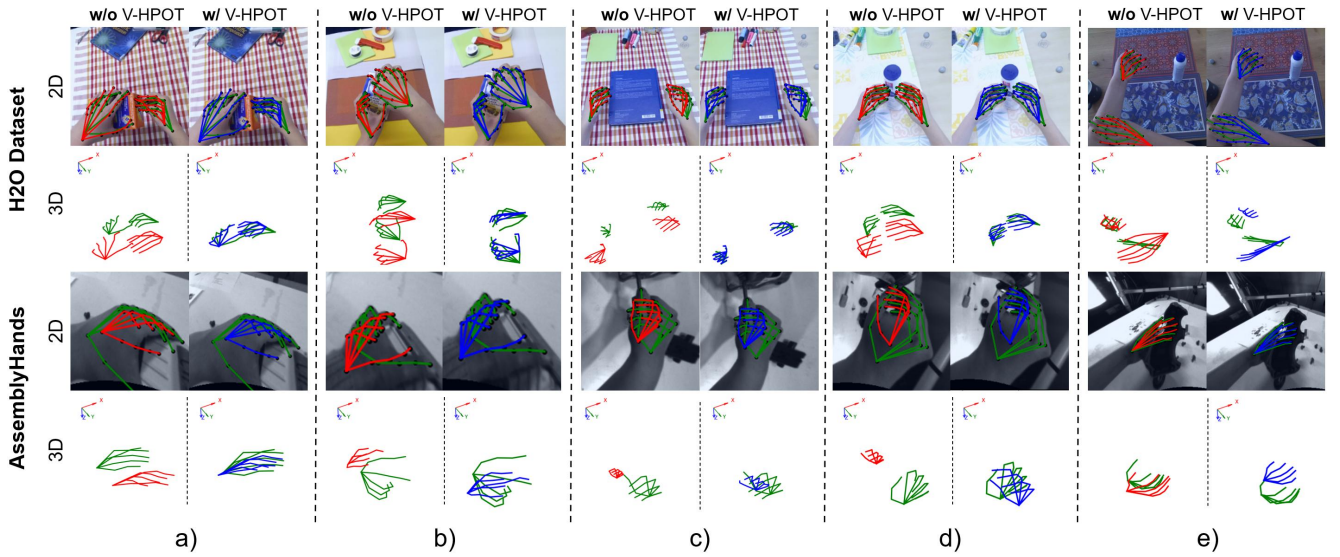


Figure 4. **Qualitative results of our method in 2D and 3D space.** Green skeletons represent the **ground truth hand pose**, red estimations **without V-HPOT** and blue estimations **with V-HPOT**. Four examples from the left (a-d) show that V-HPOT improves 3D pose estimation, while two examples from the right side (e) present a negative case with an increase of error.

5. Conclusion

We presented V-HPOT, a novel approach to egocentric 3D hand pose estimation that addresses the challenge of cross-domain generalisation. Our method introduces two key innovations: (1) a virtual camera transformation that decouples pose estimation from camera-specific intrinsics, enhancing depth representation learning; and (2) a self-supervised TTO framework that leverages consistency constraints between depth-augmented predictions. Our TTO strategy – the first to specifically target depth understanding in hand pose estimation – performs continuous weight updates during inference on the initial samples of target data. Comparison with the foundation model, particularly DINOv2, justifies the need for dedicated cross-domain hand pose studies like V-HPOT. V-HPOT’s effectiveness is confirmed through improvements in 3D absolute error (MPJPE and MRRPE) across datasets, reducing MPJPE

by 70% on *H2O*, 41% on *AssemblyHands*, despite the latter’s domain shift encompassing different scenes, sensors, and image modality (monochromatic vs. RGB). V-HPOT outperforms state-of-the-art single-stage approaches in cross-domain egocentric hand pose estimation while remaining competitive with two-stage methods despite using $\approx \times 3.5 - \times 14$ fewer training samples. The model demonstrates strong performance on real-world benchmark *Epic-Kpts* with 30% L2 error improvements, which is important as most 3D hand pose datasets are recorded in lab environments due to labelling procedures. Qualitative analysis confirms that V-HPOT produces anatomically plausible hand estimates even during complex object manipulation tasks. All this is achieved without relying on additional large-scale pretraining, datasets, or labels, confirming V-HPOT as a step forward in understanding cross-domain single-camera egocentric 3D hand pose.

Acknowledgments. This research was supported by VisuAAL ITN H2020 (grant agreement no. 861091) and by the iMars project MSCA (grant agreement no.101182996).

References

- [1] 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, 2, 5
- [2] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. OMNI3D: A Large Benchmark and Model for 3D Object Detection in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13154–13164, 2023. 2, 4, 5
- [3] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing Hand-Object Interactions in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 3, 7
- [4] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. Temporal-Aware Self-Supervised Learning for 3D Hand Pose and Mesh Estimation in Videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1050–1059, 2021. 3
- [5] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3D Hand Reconstruction via Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021. 3
- [6] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeon 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. 2, 3, 5, 7, 8
- [7] 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. 1, 5
- [8] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentation and Object Relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 5
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. 2
- [10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2, 8
- [11] Irving Fang, Yuzhong Chen, Yifan Wang, Jianghan Zhang, Qiushi Zhang, Jiali Xu, Xibo He, Weibo Gao, Hao Su, Yiming Li, et al. EgoPAT3Dv2: Predicting 3D Action Target from 2D Egocentric Vision for Human-Robot Interaction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3036–3043. IEEE, 2024. 1
- [12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1
- [13] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kaleyvtykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 2, 3
- [14] 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. 3, 7
- [15] Shengxiang Hu, Huaijiang Sun, Bin Li, Dong Wei, Weiqing Li, and Jianfeng Lu. Fast Adaptation for Human Pose Estimation via Meta-Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2024. 3, 7
- [16] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning Category-Specific Mesh Reconstruction from Image Collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2
- [17] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning Implicit Representations for Human Grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2, 3
- [18] 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. 2, 3, 5, 6, 7
- [19] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting Batch Normalization For Practical Domain Adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 3
- [20] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, and Xiaolong Wang. Test-Time Personalization

- with a Transformer for Human Pose Estimation. *Advances in Neural Information Processing Systems*, 34:2583–2597, 2021. 3, 7
- [21] Nie Lin, Takehiko Ohkawa, Yifei Huang, Mingfang Zhang, Minjie Cai, Ming Li, Ryosuke Furuta, and Yoichi Sato. SiM-Hand: Mining Similar Hands for Large-Scale 3D Hand Pose Pre-training. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2
- [23] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong 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*, pages 14687–14697, 2021. 2
- [24] Gyeongsik Moon, Shou-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. 3
- [25] Wiktor Mucha and Martin Kampel. In My Perspective, in My Hands: Accurate Egocentric 2D Hand Pose and Action Recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9, 2024. 3
- [26] Wiktor Mucha, Florin Cuconasu, Naome A. Etori, Valia Kalokyri, and Giovanni Trappolini. TEXT2TASTE: A Versatile Egocentric Vision System for Intelligent Reading Assistance Using Large Language Model. In *Computers Helping People with Special Needs*, pages 285–291, Cham, 2024. Springer Nature Switzerland. 1
- [27] Wiktor Mucha, Michael Wray, and Martin Kampel. SHARP: Segmentation of Hands and Arms by Range Using Pseudodepth for Enhanced Egocentric 3D Hand Pose Estimation and Action Recognition. In *International Conference on Pattern Recognition*, pages 178–193. Springer, 2025. 2, 3, 5, 8
- [28] 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. 2, 3, 5, 6, 7
- [29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. 7
- [30] Joonkyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 2
- [31] 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. 2, 3, 5, 8
- [32] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free Domain Adaptive Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4826–4836, 2023. 3
- [33] 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. 2, 3, 5, 7, 8
- [34] 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. 2, 4
- [35] 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. 2, 5, 8
- [36] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-Supervised 3D Hand Pose Estimation from monocular RGB via Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11230–11239, 2021. 3, 7
- [37] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020. 3
- [38] Mingxing Tan and Quoc Le. EfficientNetV2: Smaller Models and Faster Training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 3
- [39] 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. 2, 3
- [40] 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. 2, 3
- [41] Linlin Yang, Shicheng Chen, and Angela Yao. Semi-Supervised Hand Pose Estimation With Consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11364–11373, 2021. 3
- [42] 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. 6
- [43] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What’s in your hands? 3D Reconstruction of Generic Objects in Hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022. 2, 3

- [44] Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. HaMuCo: Hand Pose Estimation via Multi-view Collaborative Self-Supervised Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20763–20773, 2023. [3](#)