Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction

Yana Hasson[1,2] * Bugra Tekin[4] Federica Bogo[4]
Ivan Laptev[1,2] Marc Pollefeys[3,4] Cordelia Schmid[1,5]

[1]Inria, [2]Département d'informatique de l'ENS, CNRS, PSL Research University
[3]ETH Zürich, [4]Microsoft, [5]Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK

Abstract

Modeling hand-object manipulations is essential for understanding how humans interact with their environment. While of practical importance, estimating the pose of hands and objects during interactions is challenging due to the large mutual occlusions that occur during manipulation. Recent efforts have been directed towards fully-supervised methods that require large amounts of labeled training samples. Collecting 3D ground-truth data for hand-object interactions, however, is costly, tedious, and error-prone. To overcome this challenge we present a method to leverage photometric consistency across time when annotations are only available for a sparse subset of frames in a video. Our model is trained end-to-end on color images to jointly reconstruct hands and objects in 3D by inferring their poses. Given our estimated reconstructions, we differentiably render the optical flow between pairs of adjacent images and use it within the network to warp one frame to another. We then apply a self-supervised photometric loss that relies on the visual consistency between nearby images. We achieve state-of-the-art results on 3D hand-object reconstruction benchmarks and demonstrate that our approach allows us to improve the pose estimation accuracy by leveraging information from neighboring frames in low-data regimes.

1. Introduction

Understanding how hands interact with objects is crucial for a semantically meaningful interpretation of human action and behavior. In recent years, impressive hand pose estimation results have been demonstrated, but joint prediction of hand and object poses has received so far only limited attention, although unified 3D modeling of hands and objects is essential for many applications in augmented reality, robotics and surveillance.

*This work was performed during an internship at Microsoft.

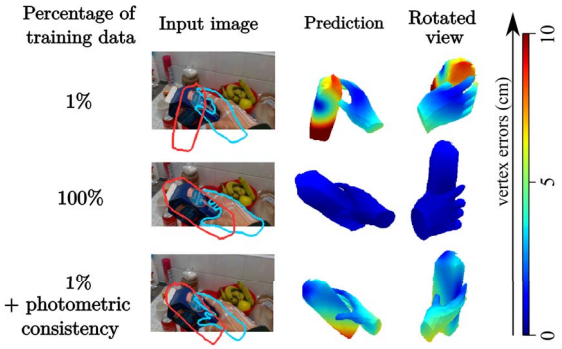

Figure 1. Our method provides accurate 3D hand-object reconstructions from monocular, sparsely annotated RGB videos. We introduce a loss which exploits photometric consistency between neighboring frames. The loss effectively propagates information from a few annotated frames to the rest of the video.

Estimating the pose of hands during interaction with an object is an extremely challenging problem due to mutual occlusions. Joint 3D reconstruction of hands and objects is even more challenging as this would require precise understanding of the subtle interactions that take place in cluttered real-world environments. Recent work in computer vision has been able to tackle some of the challenges in unified understanding of hands and objects for color input. Pioneering works of [16, 24, 37] have proposed ways to recover hand motion during object manipulation, yet without explicitly reasoning about the object pose. Recent few efforts to model hand-object interactions [12, 43], on the other hand, have focused on joint 3D hand-object pose estimation and reconstruction techniques. However, these methods require full-supervision on large datasets with 3D hand-object pose annotations. Collecting such 3D ground-truth datasets for hand-object interactions remains a challenging problem.

While motion capture datasets [6] can provide large amounts of training samples with accurate annotations, they

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00065

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

can only be captured in controlled settings and have visible markers on the images that bias pose prediction in color images. Multi-view setups [39, 52], which enable 3D triangulation from 2D detections, can similarly only be captured in controlled environments. Synthetic datasets provide an alternative. However, existing ones [12, 24, 26, 51] cannot yet reach the fidelity and realism to generalize to real datasets. Manual annotation and optimization-based techniques for data annotation can be slow and error-prone. Due to these challenges associated with data collection, existing datasets are either real ones that are limited in size and confined to constrained environments or synthetic ones that lack realism. Models trained on such data are more prone to overfitting and lack generalization capabilities.

Our method aims at tackling these challenges and reduces the stringent reliance on 3D annotations. To this end, we propose a novel weakly supervised approach to joint 3D hand-object reconstruction. Our model jointly estimates the hand and object pose and reconstructs their shape in 3D, given training videos with annotations in only sparse frames on a small fraction of the dataset. Our method models the temporal nature of 3D hand and object interactions and leverages motion as a self-supervisory signal for 3D dense hand-object reconstruction. An example result is shown in Fig. 1.

Our contributions can be summarized as follows:

- We present a new method for joint dense reconstruction of hands and objects in 3D. Our method operates on color images and efficiently regresses model-based shape and pose parameters in a single feed-forward pass through a neural network.
- We introduce a novel photometric loss that relies on the estimated optical flow between pairs of adjacent images. Our scheme leverages optical flow to warp one frame to the next, directly within the network, and exploits the visual consistency between successive warped images with a self-supervised loss, ultimately alleviating the need for strong supervision.

In Section 4, we show quantitatively that these contributions allow us to reliably predict the pose of interacting hands and objects in 3D, while densely reconstructing their 3D shape. Our approach allows us to improve pose estimation accuracy in the absence of strong supervision on challenging real-world sequences and achieves state-of-the-art results on 3D hand-object reconstruction benchmarks. The code is publicly available. [1]

2. Related Work

Our work tackles the problem of estimating hand-object pose from monocular RGB videos, exploiting photometric cues for self-supervision. To the best of our knowledge, our method is the first to apply such self-supervision to hand-object scenarios. We first review the literature on hand and object pose estimation. Then, we focus on methods using motion and photometric cues for self-supervision, in particular in the context of human body pose estimation.

Hand and object pose estimation. Most approaches in the literature tackle the problem of estimating either hand or object pose, separately.

For object pose estimation from RGB images, the recent trend is to use convolutional neural networks (CNNs) to predict the 2D locations of the object's 3D bounding box in image space [20, 33, 44]. The 6D pose is then obtained via PnP [21] or further iterative refinement. Such methods commonly need a 3D model of the object as input, and large amounts of labeled data. DeepIM [22] shows generalization to unseen objects by iteratively matching rendered images of an object against observed ones. Recently, Pix2Pose [28] improves robustness against occlusions by predicting dense 2D-3D correspondences between image pixels and the object model. Most methods [22, 28, 42] try to limit the amount of required annotations by relying on synthetic data. However, it remains unclear how well these methods would perform in the presence of large occlusions as the ones caused by hand-object interactions.

Several approaches for hand pose estimation from RGB images focus on regressing 3D skeleton joint positions [5, 16, 24, 40, 48, 51]. However, methods that directly output 3D hand surfaces offer a richer representation, and allow one to directly reason about occlusions and contact points [25]. Parametric hand models like MANO [38] represent realistic 3D hand meshes using a set of shape and pose parameters. [17, 30] fit such parametric models to CNN-regressed 2D joint positions to estimate hand poses from full-body images.

A number of recent methods plug MANO into end-to-end deep learning frameworks, obtaining accurate hand 3D shape and pose from single RGB images [3, 7, 49]. Similarly to us, these approaches regress the model parameters directly from the image, though they do not address scenarios with hand-object interactions. Given the challenges involved, hand-object interactions have been tackled in multi-view or RGB-D camera setups [36, 46]. Targeting pose estimation from single RGB images, Romero et al. [37] obtain 3D hand-object reconstructions via nearest neighbor search in a large database of synthetic images.

Recently, efforts have been put into the acquisition of ground-truth 3D annotations for both hands and objects during interaction. Early datasets which provide annotated RGB views of hands manipulating objects rely on manual annotations [41] and depth tracking [46], which limits the size and the occlusions between hand and object. Larger

[1] https://hassony2.github.io/handobjectconsist

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

datasets which rely on motion capture [6] and multi-view setups [11] have been collected, spurring the development of new methods for hand-object pose estimation. Recently, [12, 43] propose CNN-based approaches to accurately predict hand and object poses from monocular RGB. However, these methods are fully supervised and do not exploit the temporal dimension for pose estimation.

Supervision using motion and photometric cues. In RGB videos, motion cues provide useful information that can be used for self-supervision. Several methods explore this idea in the context of human body pose estimation.

Pfister et al. [32] leverage optical flow for 2D human pose estimation. Slim DensePose [27] uses an off-the-shelf optical flow method [15] to establish dense correspondence [9] between adjacent frames in a video. These correspondences are used to propagate manual annotations between frames and to enforce spatio-temporal equivariance constraints. Very recently, PoseWarper [2] leverages image features to learn the pose warping between a labeled frame and an unlabeled one, thus propagating annotations in sparsely labeled videos.

Regressing 3D poses is more difficult: the problem is fundamentally ambiguous in monocular scenarios. Furthermore, collecting 3D annotations is not as easy as in 2D. VideoPose3D [31] regresses 3D skeleton joint positions, by back-projecting them on the image space and using CNN-estimated 2D keypoints as supervision. Tung et al. [45] regress the SMPL body model parameters [23] by using optical flow and reprojected masks as weak supervision. Differently from us, they rely on an off-the-shelf optical flow method, making the pose accuracy dependent on the flow quality. Recently, Arnab et al. [1] refine noisy per-frame pose predictions [18] using bundle adjustment over the SMPL parameters. These methods are not tested in scenarios with large body occlusions.

Our method enforces photometric consistency between pose estimates from adjacent frames. Similar ideas have been successfully applied to self-supervised learning of ego-motion, depth and scene flow for self-driving cars [4, 8, 50]. Unlike these methods, which estimate pixel-wise probability depth distributions for mostly rigid scenes, we focus on estimating the articulated pose of hands manipulating objects. Starting from multi-view setups at training time, [34, 35] propose weak supervision strategies for monocular human pose estimation. We consider monocular setups where the camera might move. Similarly to us, TexturePose [29] enforces photometric consistency between pairs of frames to refine body pose estimates. They define the consistency loss in UV space: this assumes a UV parameterization is always provided. Instead, we define our loss in image space. Notably, these methods consider scenarios without severe occlusions (only one instance, *i.e.* one body, is in the scene).

None of these methods focuses on hands, and more particularly on complex hand-object interactions.

3. Method

We propose a CNN-based model for 3D hand-object reconstruction that can be efficiently trained from a set of *sparsely annotated* video frames. Namely, our method takes as input a monocular RGB video, capturing hands interacting with objects. We assume that the object model is known, and that sparse annotations are available only for a subset of video frames.

As in previous work [20, 43, 44], we assume that a 3D mesh model of the object is provided. To reconstruct hands, we rely on the parametric model MANO [38], which deforms a 3D hand mesh template according to a set of shape and pose parameters. As output, our method returns hand and object 3D vertex locations (together with shape and pose parameters) for each frame in the sequence.

The key idea of our approach is to use a photometric consistency loss, that we leverage as self-supervision on the unannotated intermediate frames in order to improve hand-object reconstructions. We introduce this loss in Sec. 3.1. We then describe our learning framework in detail in Sec. 3.2.

3.1. Photometric Supervision from Motion

As mentioned above, our method takes as input a sequence of RGB frames and outputs hand and object mesh vertex locations for each frame. The same type of output is generated in [12], where each RGB frame is processed separately. We observe that the temporal continuity in videos imposes temporal constraints between neighboring frames. We assume that 3D annotations are provided only for a sparse subset of frames; this is a scenario that often occurs in practice when data collection is performed on sequential images, but only a subset of them is manually annotated. We then define a self-supervised loss to propagate this information to unlabeled frames.

Our self-supervised loss exploits photometric consistency between frames, and is defined in image space. Figure 2 illustrates the process. Consider an annotated frame at time t_{ref}, $I_{t_{ref}}$, for which we have ground-truth hand and object vertices $V_{t_{ref}}$ (to simplify the notation, we do not distinguish here between hand and object vertices). Given an unlabeled frame $I_{t_{ref}+k}$, our goal is to accurately regress hand and object vertex locations $V_{t_{ref}+k}$. Our main insight is that, given estimated per-frame 3D meshes and known camera intrinsics, we can back-project our meshes on image space and leverage pixel-level information to provide additional cross-frame supervision.

Given $I_{t_{ref}+k}$, we first regress hand and object vertices $V_{t_{ref}+k}$ in a single feed-forward network pass (see

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

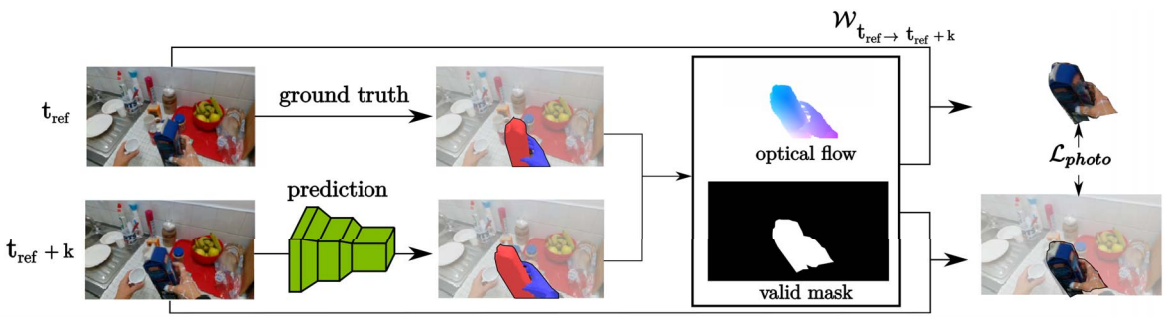

Figure 2. Photometric consistency loss. Given an annotated frame, t_{ref}, and an unannotated one, t_{ref+k}, we reconstruct hand and object 3D pose at t_{ref+k} leveraging a self-supervised loss. We differentiably render the optical flow between ground-truth hand-object vertices at t_{ref} and estimated ones. Then, we use this flow to warp frame t_{ref+k} into t_{ref}, and enforce consistency in pixel space between warped and real image.

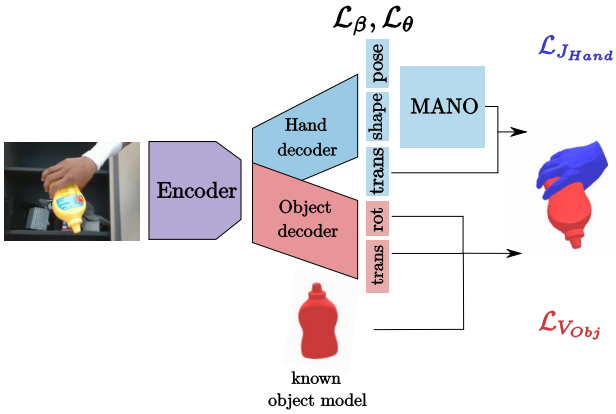

Figure 3. Architecture of the single-frame hand-object reconstruction network.

Sec. 3.2). Imagine now to back-project these vertices on $I_{t_{ref}+k}$ and assign to each vertex the color of the pixel they are projected onto. The object meshes at t_{ref} and t_{ref+k} share the same topology; and so do the hand meshes. So, if we back-project the ground-truth meshes at t_{ref} on $I_{t_{ref}}$, corresponding vertices from $V_{t_{ref}}$ and $V_{t_{ref}+k}$ should be assigned the same color.

We translate this idea into our photometric consistency loss. We compute the 3D displacement ("flow") between corresponding vertices from $V_{t_{ref}}$ and $V_{t_{ref}+k}$. These values are then projected on the image plane, and interpolated on the visible mesh triangles. To this end, we differentiably render the estimated flow from $V_{t_{ref}}$ to $V_{t_{ref}+k}$ using the Neural Renderer [19]. This allows us to define a warping flow W between the pair of images as a function of $V_{t_{ref}+k}$.

We exploit the computed flow to warp $I_{t_{ref}+k}$ into the warped image $\mathcal{W}(I_{t_{ref}+k}, V_{t_{ref}+k})$, by differentiably sampling values from $I_{t_{ref}+k}$ according to the predicted optical flow displacements. Our loss enforces consistency between the warped image and the reference one:

$$\mathcal{L}_{photo}(V_{t_{ref}+k}) = ||M \cdot (\mathcal{W}(I_{t_{ref}+k}, V_{t_{ref}+k}) - I_{t_{ref}})||_1, \quad (1)$$

where M is a binary mask denoting surface point visibility. In order to compute the visibility mask, we ensure that the supervised pixels belong to the silhouette of the reprojected mesh in the target frame $I_{t_{ref}+k}$. We additionally verify that the supervision is not applied to pixels which are occluded in the reference frame by performing a cyclic consistency check similarly to [14, 27] which is detailed in the supplemental material. We successively warp a grid of pixel locations using the optical flow t_{ref} to $t_{ref}+k$ and from $t_{ref}+k$ to t_{ref} and include only pixel locations which remain stable, a constraint which does not hold for mesh surface points which are occluded in one of the frames. Note that the error is minimized with respect to the estimated hand and object vertices $V_{t_{ref}+k}$.

The consistency supervision $\mathcal{L}_{photo}$ can be applied directly on pixels, similarly to self-supervised ego-motion and depth learning scenarios [8, 50]. The main difference with these approaches is that they estimate per-pixel depth values while we attempt to leverage the photometric consistency loss in order to refine rigid and articulated motions. Our approach is similar in spirit to that of [29]. With respect to them, we consider a more challenging scenario (multiple 3D instances and large occlusions). Furthermore, we define our loss in image space, instead of UV space, and thus we do not assume that a UV parametrization is available.

As each operation is differentiable, we can combine this loss and use it as supervision either in isolation or in addition to other reconstruction losses (Sec. 3.2).

3.2. Dense 3D Hand-Object Reconstruction

We apply the loss introduced in Sec. 3.1 to 3D hand-object reconstructions obtained independently for each frame. These per-frame estimates are obtained with a single forward pass through a deep neural network, whose architecture is shown in Fig. 3. In the spirit of [3, 12], our network takes as input a single RGB image and outputs MANO [38] pose and shape parameters. However, differently from [12], we assume that a 3D model of the object is given, and we regress its 6D pose by adding a second head to our network (see again Fig. 3). We employ as backbone

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

a simple ResNet-18 [13], which is computationally very efficient (see Sec. 4). We use the base network model as the image encoder and select the last layer before the classifier to produce our image features. We then regress hand and object parameters from these features through 2 dense layers with ReLU non-linearities. Further details about the architecture can be found in the supplemental material.

In the following, we provide more details about hand-object pose and shape regression, and about the losses used at training time.

Hand-object global pose estimation. We formulate the hand-object global pose estimation problem in the camera coordinate system and aim to find precise absolute 3D positions of hands and objects. Instead of a weak perspective camera model, commonly used in the body pose estimation literature, we choose here to use a more realistic projective model. In our images, hand-object interactions are usually captured at a short distance from the camera. So the assumptions underlying weak perspective models do not hold. Instead, we follow best practices from object pose estimation. As in [22, 47], we predict values that can be easily estimated from image evidence. Namely, in order to estimate hand and object translation, we regress a focal-normalized depth offset d_f and a 2D translation vector (t_u, t_v), defined in pixel space. We compute d_f as

$$d_f = \frac{V_z - z_{off}}{f}, \tag{2}$$

where V_z is the distance between mesh vertex and camera center along the z-axis, f is the camera focal length, and z_{off} is empirically set to $40cm$. t_u and t_v represent the translation, in pixels, of the object (or hand) origin, projected on the image space, with respect to the image center. Note that we regress d_f and (t_u, t_v) for both the hand and the object, separately.

Given the estimated d_f and (t_u, t_v), and the camera intrinsics parameters, we can easily derive the object (hand) global translation in 3D. For the global rotation, we adopt the axis-angle representation. Following [18, 22, 30], the rotation for object and hand is predicted in the object-centered coordinate system.

Hand articulated pose and shape estimation. We obtain hand 3D reconstructions by predicting MANO pose and shape parameters. For the pose, similarly to [3, 12], we predict the principal composant analysis (PCA) coefficients of the low-dimensional hand pose space provided in [38]. For the shape, we predict the MANO shape parameters, which control identity-specific characteristics such as skeleton bone length. Overall, we predict 15 pose coefficients and 10 shape parameters.

Regularization losses. We find it effective to regularize both hand pose and shape by applying ℓ_2 penalization as in [3]. $\mathcal{L}_{\theta_{Hand}}$ prevents unnatural joint rotations, while $\mathcal{L}_{\beta_{Hand}}$ prevents extreme shape deformations, which can result in irregular and unrealistic hand meshes.

Skeleton adaptation. Hand skeleton models can vary substantially between datasets, resulting in inconsistencies in the definition of joint locations. Skeleton mismatches may force unnatural deformations of the hand model. To account for these differences, we replace the fixed MANO joint regressor with a skeleton adaptation layer which regresses joint locations from vertex positions. We initialize this linear regressor using the values from the MANO joint regressor and optimize it jointly with the network weights. We keep the tips of the fingers and the wrist joint fixed to the original locations, and learn a dataset-specific mapping for the other joints at training time.

Reconstruction losses. In total, we predict 6 parameters for hand-object rotation and translation and 25 MANO parameters, which result in a total of 37 regressed parameters. We then apply the predicted transformations to the reference hand and object models and further produce the 3D joint locations of the MANO hand model, which are output by MANO in addition to the hand vertex locations. We define our supervision on hand joint positions, $\mathcal{L}_{J_{Hand}}$, as well as on 3D object vertices, $\mathcal{L}_{V_{Obj}}$. Both losses are defined as ℓ_2 errors.

Our final loss $\mathcal{L}_{HO}$ is a weighted sum of the reconstruction and regularization terms:

$$\mathcal{L}_{HO} = \mathcal{L}_{V_{Obj}} + \lambda_J \mathcal{L}_{J_{Hand}} + \lambda_\beta \mathcal{L}_{\beta_{Hand}} + \lambda_\theta \mathcal{L}_{\theta_{Hand}}. \tag{3}$$

4. Evaluation

In this section, we first describe the datasets and corresponding evaluation protocols. We then compare our method to the state of the art and provide a detailed analysis of our framework.

4.1. Datasets

We evaluate our framework for joint 3D hand-object reconstruction and pose estimation on two recently released datasets: First Person Hand Action Benchmark [6] and HO-3D [11] which provide pose annotations for all hand keypoints as well as the manipulated rigid object.

First-person hand action benchmark (FPHAB): The FPHAB dataset [6] collects egocentric RGB-D videos capturing a wide range of hand-object interactions, with ground-truth annotations for 3D hand pose, 6D object pose, and hand joint locations. The annotations are obtained in an automated way, using mocap magnetic sensors strapped on hands. Object pose annotations are available for 4 objects, for a subset of the videos. Similarly to hand annotations, they are obtained via magnetic sensors. In our evaluation, we use the same *action split* as in [43]: each object is present in both the training and test splits, thus allowing the model to learn instance-specific 6 degrees of freedom

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

(DoF) transformations. To further compare our results to those of [12], we also use the *subject split* of FPHAB where the training and test splits feature different subjects.

HO-3D: The recent HO-3D dataset [11] is the result of an effort to collect 3D pose annotations for both hands and manipulated objects in a markerless setting. In this work, we report results on the subset of the dataset which was released as the first version [10]. Details on the specific subset are provided in supplemental material. The subset of HO-3D we focus on contains 14 sequences, out of which 2 are available for evaluation. The authors augment the real training sequences with additional synthetic data. In order to compare our method against the baselines introduced in [10], we train jointly on their real and synthetic training sets.

4.2. Evaluation Metrics

We evaluate our approach on 3D hand pose estimation and 6D object pose estimation and use official train/test splits to evaluate our performance in comparison to the state of the art. We report accuracy using the following metrics.

Mean 3D errors. To assess the quality of our 3D hand reconstructions, we compute the mean end-point error (in mm) over 21 joints following [51]. For objects, on FPHAB we compute the average vertex distance (in mm) in camera coordinates to compare against [43], on HO-3D, we look at average bounding box corner distances.

Mean 2D errors. We report the mean errors between reprojected keypoints and 2D ground-truth locations for hands and objects. To evaluate hand pose estimation accuracy, we measure the average joint distance. For object pose estimation, following the protocol for 3D error metrics, we report average 2D vertex distance on FPHAB, and average 2D corner distance on HO-3D. To further compare our results against [10], we also report the percentage of correct keypoints (PCK). To do so, for different pixel distances, we compute the percentage of frames for which the average error is lower than the given threshold.

4.3. Experimental Results

We first report the pose estimation accuracy of our single-frame hand-object reconstruction model and compare it against the state of the art [10, 43]. We then present the results of our motion-based self-supervised learning approach and demonstrate its efficiency in case of scarcity of ground-truth annotations.

Single-frame hand-object reconstruction. Taking color images as input, our model reconstructs dense meshes to leverage pixel-level consistency, and infers hand and object poses. To compare our results to the state of the art [10, 12, 43], we evaluate our pose estimation accuracy on the FPHAB [6] and HO-3D [11] datasets.

Method	Hand error	Object error
Tekin *et al.*	**15.8**	24.9
Ours	18.0	**22.3**

Table 1. Comparison to state-of-the-art method of Tekin *et al.* [43] on FPHAB [6], errors are reported in mm.

Method	Hand error
Ours - no skeleton adaptation	28.1
Ours	**27.4**
Hasson *et al.* [12]	28.0

Table 2. On the FHPAB dataset, for which the skeleton is substantially different from the MANO one, we show that adding a skeleton adaptation layer allows us to outperform [12], while additionally predicting the global translation of the hand.

Table 1 demonstrates that our model achieves better accuracy than [43] on object pose estimation. We attribute this to the fact that [43] regresses keypoint positions, and recovers the object pose as a non-differentiable post-processing step, while we directly optimize for the 6D pose. Our method achieves on average a hand pose estimation error of 18 mm on FPHAB which is outperformed by [43] by a margin of 2.6 mm. This experiment is in line with earlier reported results, where the estimation of individual keypoint locations outperformed regression of model parameters [18, 29, 30]. While providing competitive pose estimation accuracy to the state of the art, our approach has the advantage of predicting a detailed hand shape, which is crucial for fine-grained understanding of hand-object interactions and contact points. We further compare our results to those of [12] that reports results on FPHAB using the *subject split* and demonstrate that our model provides improved hand pose estimation accuracy, while additionally estimating the global position of the hand in the camera space.

We further evaluate the hand-object pose estimation accuracy of our single-image model on the recently introduced HO-3D dataset. We show in Fig. 5 that we outperform [10] on both hand and object pose estimation.

In Table 3, we analyze the effect of simultaneously training for hand and object pose estimation within a unified framework. We compare the results of our unified model to those of the models trained individually for hand pose estimation and object pose estimation. We observe that the unified co-training slightly degrades hand pose accuracy. This phenomenon is also observed by [43], and might be due to the fact that while the hand pose highly constrains the object pose, simultaneous estimation of the object pose does not result in increased hand pose estimation accuracy, due to higher degrees of freedom inherent to the articulated pose estimation problem.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

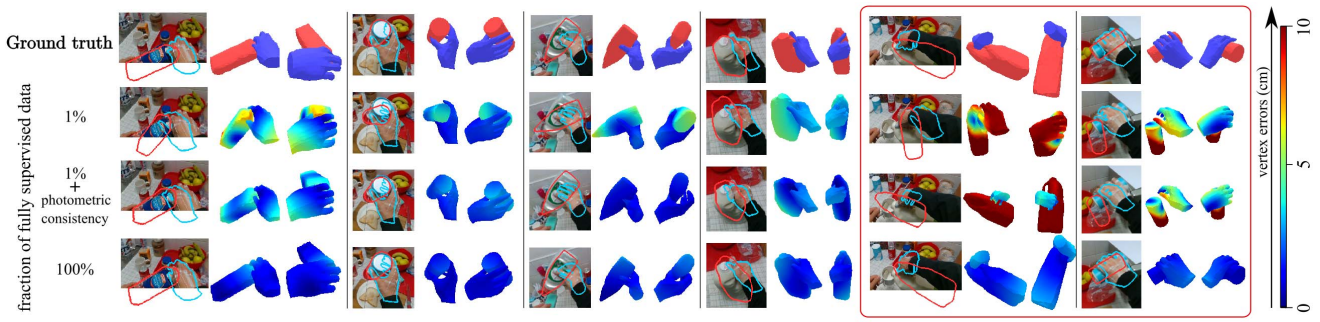

Figure 4. Qualitative results on the FPHAB dataset. We visualize the reconstructed meshes reprojected on the image as well as a rotated view. When training on the full dataset, we obtain reconstructions which accurately capture the hand-object interaction. In the sparsely supervised setting, we qualitatively observe that photometric consistency allows to recover more accurate hand and object poses. Failure cases occur in the presence of important motion blur and large occlusions of the hand or the object by the subject's arm.

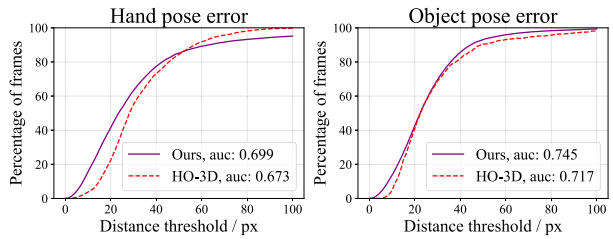

Figure 5. Evaluation of our baseline for hand-object pose estimation on HO-3D. We report the PCK for 2D joint mean-end-point error for hands, and the mean 2D reprojection error for objects.

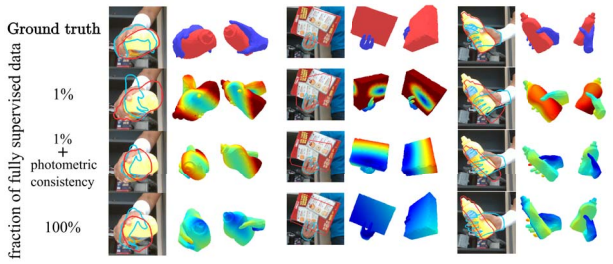

Figure 6. Predicted reconstructions for images from HO-3D. While rotation errors around axis parallel to the camera plane are not corrected and are sometimes even introduced by the photometric consistency loss, we observe qualitative improvement in the 2D reprojection of the predicted meshes on the image plane.

	Hand error (mm)	Object error (mm)
Hand only	15.7	-
Object only	-	21.8
Hand + Object	18.0	22.3

Table 3. We compare training for hand and object pose estimation jointly and separately on FPHAB [6] and find that the encoder can be shared at a minor performance cost in hand and object pose accuracy.

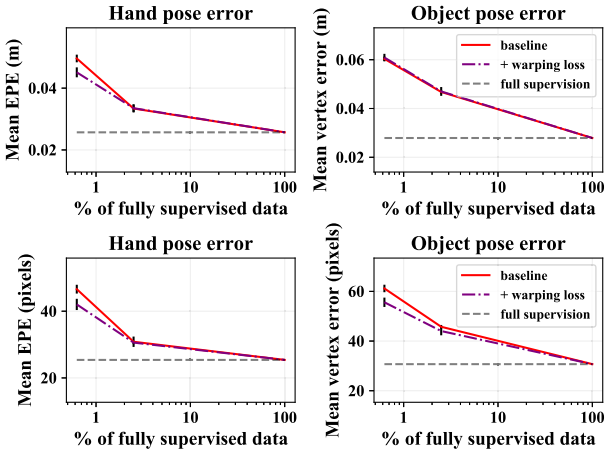

Figure 7. Effect of using photometric-consistency self-supervision when only a fraction of frames are fully annotated on HO-3D. We report average values and standard deviations over 5 different runs.

Photometric supervision on video. We now validate the efficiency of our self-supervised dense hand-object reconstruction approach when ground-truth data availability is limited. We pretrain several models on a fraction of the data by sampling frames uniformly in each sequence. We sample a number of frames to reach the desired ratio of annotated frames in each training video sequence, starting from the first frame. We then continue training with photometric consistency as an additional loss, while maintaining the full supervision on the sparsely annotated frames. Additional implementation and training details are discussed in supplemental material. In order to single out the effect of the additional consistency term and factor out potential benefits from a longer training time, we continue training a reference model with the full supervision on the sparse keyframes for comparison. We experiment with various

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

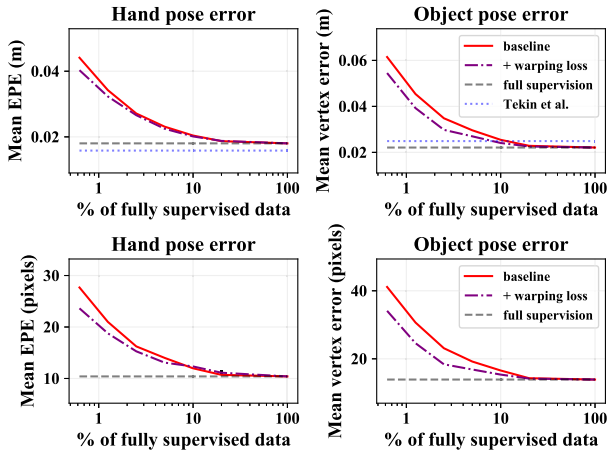

Figure 8. We observe consistent quantitative improvements from the photometric consistency loss as the percentage of fully supervised frames decreases below 10% for both hands and objects.

regimes of data scarcity, progressively decreasing the percentage of annotated keyframes from 50 to less than 1%.

We report our results in Fig. 8 for FPHAB and in Fig. 7 for HO-3D. We observe that only 20% of the frames are necessary to reach the densely supervised performance on the FPHAB dataset, which can be explained by the correlated nature between neighboring frames. However, as we further decrease the fraction of annotated data, the generalization error significantly decreases. We demonstrate that our self-supervised learning strategy significantly improves the pose estimation accuracy in the low data regime when only a few percent of the actual dataset size are annotated and reduces the rigid reliance on large labeled datasets for hand-object reconstruction. Although the similarity between the reference and consistency-supervised frames decreases as the supervision across video becomes more sparse and the average distance to the reference frame increases, resulting in larger appearance changes, we observe that the benefits from our additional photometric consistency is most noticeable for both hands and objects as scarcity of fully annotated data increases. When using less than one percent of the training data with full supervision, we observe an absolute average improvement of 7 pixels for objects and 4 pixels for hands, reducing the gap between the sparsely and fully supervised setting by respectively 25 and 23% (see Fig. 8). While on HO-3D the pixel-level improvements on objects do not translate to better 3D reconstruction scores for the object (see Fig. 7), on FPHAB, the highest relative improvement is observed for object poses when fully supervising 2.5% of the data. In this setup, the 4.7 reduction in the average pixel error corresponds to a reduction of the error by 51% and results in a reduction by 40% in the 3D mm error. We qualitatively investigate the modes

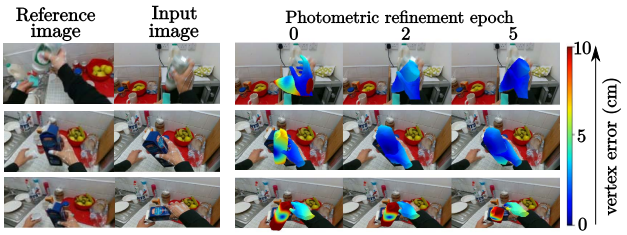

Figure 9. Progressive pose refinement over training samples, even in the presence of large motion and inaccurate initialization. In extreme cases (last row), the model cannot recover.

of improvement and failure from introducing the additional photometric consistency loss in Fig. 4 and Fig. 6.

As our method relies on photometric consistency for supervision, it is susceptible to fail when the photometric consistency assumption is infringed, which can occur for instance in cases of fast motions or illumination changes. However, our method has the potential to provide meaningful supervision in cases where large motions occur between the reference and target frames, as long as the photometric consistency hypothesis holds. We observe that in most cases, our baseline provides reasonable initial pose estimates on unannotated frames, which allows the photometric loss to provide informative gradients. In Fig. 9, we show examples of successful and failed pose refinements on training samples from the FPHAB dataset supervised by our loss. Our model is able to improve pose estimations in challenging cases, where the initial prediction is inaccurate and there are large motions with respect to the reference frame.

5. Conclusion

In this paper, we propose a new method for dense 3D reconstruction of hands and objects from monocular color images. We further present a sparsely supervised learning approach leveraging photo-consistency between sparsely supervised frames. We demonstrated that our approach achieves high accuracy for hand and object pose estimation and successfully leverages similarities between sparsely annotated and unannotated neighboring frames to provide additional supervision. Future work will explore additional self-supervised 3D interpenetration and scene interaction constraints for hand-object reconstruction. Our framework is general and can be extended to incorporate the full 3D human body along with the environment surfaces, which we intend to explore to achieve a full human-centric scene understanding.

Acknowledgments. This work was supported by the MSR-Inria joint lab and the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). Yana Hasson thanks Mihai Dusmanu, Yann Labbé and Thomas Eboli for helpful discussions.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

References

[1] Anurag* Arnab, Carl* Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[2] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. In *Advances in Neural Information Processing Systems*, 2019. 3

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3d hand shape and pose from images in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4, 5

[4] Fabian Brickwedde, Steffen Abraham, and Rudolf Mester. Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[5] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3D hand pose estimation from monocular RGB images. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[6] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 5, 6, 7

[7] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4

[9] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[10] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. In *arXiv Preprint 1907.01481v1*, 2019. 6

[11] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and objects poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 5, 6

[12] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 4, 5, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[14] Junhwa Hur and Stefan Roth. MirrorFlow: Exploiting symmetries in joint optical flow and occlusion estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[16] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *The European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[17] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 5, 6

[19] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[20] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3

[21] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision (IJCV)*, 2009. 2

[22] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *The European Conference on Computer Vision (ECCV)*, 2018. 2, 5

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2015. 3

[24] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[25] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics (TOG)*. 2

[26] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[27] Natalia Neverova, James Thewlis, Riza Alp Guler, Iasonas Kokkinos, and Andrea Vedaldi. Slim densepose: Thrifty learning from sparse annotations and motion cues. In *The*

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 3, 4

[28] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[29] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 4, 6

[30] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6

[31] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[32] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[33] Mahdi Rad and Vincent Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[34] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation learning for 3D human pose estimation. In *The European Conference on Computer Vision (ECCV)*, 2018. 3

[35] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3D human pose estimation from multi-view images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[36] Grégory Rogez, James Steven Supancic III, and Deva Ramanan. Understanding everyday hands in action from RGB-D images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[37] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: real-time 3D reconstruction of hands in interaction with objects. 2010. 1, 2

[38] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 2, 3, 4, 5

[39] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[40] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[41] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *The European Conference on Computer Vision (ECCV)*, 2016. 2

[42] Martin Sundermeyer, Marton Zoltan-Csaba, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3D orientation learning for 6D object detection from RGB images. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[43] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 5, 6

[44] Bugra Tekin, Sudipta Sinha, and Pascal Fua. Real-time seamless single shot 6D object pose prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[45] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, 2017. 3

[46] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, 2016. 2

[47] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems (RSS)*, 2018. 5

[48] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[49] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[50] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe. Unsupervised learning of depth and ego-motion from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4

[51] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6

[52] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:22:19 UTC from IEEE Xplore. Restrictions apply.