The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Collaborative Learning for 3D Hand-Object Reconstruction and Compositional Action Recognition from Egocentric RGB Videos Using Superquadrics

Tze Ho Elden Tse[1], **Runyang Feng**[2], **Linfang Zheng**[1], **Jiho Park**[3], **Yixing Gao**[2*], **Jihie Kim**[3], **Aleš Leonardis**[1], **Hyung Jin Chang**[1]
[1]University of Birmingham
[2]Jilin University
[3]Dongguk University

Abstract

With the availability of egocentric 3D hand-object interaction datasets, there is increasing interest in developing unified models for hand-object pose estimation and action recognition. However, existing methods still struggle to recognise seen actions on unseen objects due to the limitations in representing object shape and movement using 3D bounding boxes. Additionally, the reliance on object templates at test time limits their generalisability to unseen objects. To address these challenges, we propose to leverage superquadrics as an alternative 3D object representation to bounding boxes and demonstrate their effectiveness on both template-free object reconstruction and action recognition tasks. Moreover, as we find that pure appearance-based methods can outperform the unified methods, the potential benefits from 3D geometric information remain unclear. Therefore, we study the compositionality of actions by considering a more challenging task where the training combinations of verbs and nouns do not overlap with the testing split. We extend H2O and FPHA datasets with compositional splits and design a novel collaborative learning framework that can explicitly reason about the geometric relations between hands and the manipulated object. Through extensive quantitative and qualitative evaluations, we demonstrate significant improvements over the state-of-the-arts in (compositional) action recognition.

Introduction

Understanding hand and object interaction is a fundamental problem in various downstream applications including augmented and virtual reality (AR/VR) (Han et al. 2020, 2022; Tse et al. 2023). While a significant amount of research has focused on estimating the poses of hand and objects (Tse et al. 2022a,b; Hampali et al. 2022; Chen et al. 2022; Fan et al. 2023; Chen et al. 2023; Ye et al. 2023b; Feng et al. 2023b,a), jointly recognising hand-object interactions has received far less attention. In this paper, as illustrated in Fig. 1, we study jointly the problem of hand-object reconstruction and interaction recognition from egocentric view.

Recent egocentric hand-object interaction datasets (Kwon et al. 2021; Garcia-Hernando et al. 2018) with 3D annotations enable the development of a unified framework for estimating hand-object poses and interaction classes (Tekin, Bogo, and Pollefeys 2019; Yang et al. 2020; Kwon et al. 2021; Wen et al. 2023; Cho et al. 2023). These methods couple 3D geometric cues (*i.e.* hand-object poses and/or contact maps) and appearance features to predict interaction class. Despite a unified understanding of the hand and manipulated object dynamics being crucial for recognising egocentric interactions, we find that pure appearance-based methods (*i.e.* MViT (Fan et al. 2021; Li et al. 2022)) can achieve comparable performance to the state-of-the-arts (as shown in Table 1). This raises immediate questions on when or how 3D geometric features can benefit interaction recognition.

In contrast, while deep architectures trained on large-scale datasets (Sigurdsson et al. 2016; Kay et al. 2017; Karpathy et al. 2014) exhibit strong distribution learning capabilities, mainstream action recognition models (Simonyan and Zisserman 2014; Carreira and Zisserman 2017; Feichtenhofer et al. 2019; Wang et al. 2016) primarily focus on frame appearance rather than temporal reasoning. Consequently, reversing the order of the video frame at test time will often produce the same classification result (Materzynska et al. 2020; Zhou et al. 2018). In particular, classical activity recognition methods like the two-

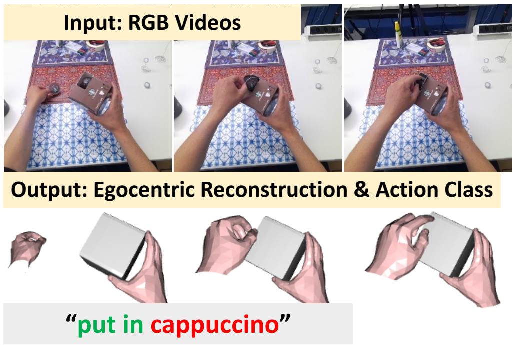

Figure 1: Our method jointly reconstructs hand-object meshes without object instance-specific templates and recognises interaction from egocentric RGB video. We further consider a more challenging problem scenario, compositional action recognition, where combinations of verb (in green) and noun (in red) are unseen during training. Our model is designed for generalising action recognition by explicitly leveraging 3D geometric information.

---

*Corresponding author.

stream Convolutional Neural Network (Simonyan and Zisserman 2014) and I3D (Carreira and Zisserman 2017) have demonstrated strong performance on various video datasets, including UCF101 (Soomro, Zamir, and Shah 2012) and Sport1M (Karpathy et al. 2014), with only still frames and optical flow. While appearance features can be highly predictive of the action class (Santoro et al. 2017; Battaglia et al. 2018), it remains challenging for appearance-based deep networks to capture the *compositionality* of action and objects without temporal transformations or geometric relations (Materzynska et al. 2020).

To address the aforementioned problems, Materzynska et al. (2020) extends the Something-Something dataset (Goyal et al. 2017) and introduces the Something-Else task with a new compositional split. This presents a novel task known as compositional action recognition, in which methods are required to recognise an action with unseen objects. Under this problem setting, the combinations of actions and object instances do not overlap in the training and testing split. Therefore, models are encouraged to learn the compositionality of action *verb* and *noun*, and not overfit to the correlation between appearance features and action classes. Nonetheless, the current research in this task is primarily generic approaches using 2D geometric cues such as 2D instance bounding boxes. The potential benefits offered by 3D geometric information remain an open problem.

Therefore, we take an alternative approach which exploits the compositionality of actions using 3D geometric information. To achieve that, we first extend the two existing 3D annotated egocentric hand-object datasets, H2O (Kwon et al. 2021) and FPHA (Garcia-Hernando et al. 2018), by introducing new compositional splits. Our experiments show that the existing state-of-the-art approaches (Yang et al. 2020; Wen et al. 2023; Fan et al. 2021; Li et al. 2022) still face significant challenges in recognising a seen action when facing new objects. This is because the current methods (either single or dual branches) are unable to tackle the problem of appearance bias in objects, as they have to take the combination of appearance and geometric information as a whole. In addition, these approaches focus on extracting features for the whole scene and do not explicitly recognise objects as individual entities. Hence, they cannot fully capture the compositionality of the action.

In this paper, we propose a collaborative learning framework that allows an action verb and object to interact and complement each other. The key motivation for this strategy is that the tasks of estimating hand-object poses and recognising interactions are naturally closely-correlated. Existing collaborative learning methods (Yang et al. 2020; Tse et al. 2022a) in understanding hand-object interactions typically follow an iterative approach where the multiple target learning tasks (*i.e.* hand-object pose estimation and action recognition) boost each other mutually and progressively. However, connecting branches iteratively can lead to highly unstable training (Tse et al. 2022a). This is because gradients from one branch can propagate through the connections to affect the other branches which causes unstable gradients. We explicitly address this by a new transformer-based design to exploit the compositionality of actions and avoid branch stacking. In addition, we propose to use superquadrics (Barr 1981) as the intermediate 3D object representation. This is motivated by the fact that existing action recognition methods have limitations in accurately representing objects' shape and movement with only 2D/3D bounding boxes. But at the same time, accurately reconstructing 3D objects without object templates remains highly challenging, especially in scenarios involving unseen objects. Therefore, superquadrics offer a compact representation with their ability to represent a wide range of shapes with few parameters. In addition, it allows models to interpret objects with basic geometric primitives.

Our contributions are the following:

1. We propose an end-to-end collaborative learning framework to leverage 3D geometric information for compositional action recognition from egocentric RGB videos.
2. We show that using superquadrics as the intermediate 3D object representation is beneficial for 3D hand pose estimation and interaction recognition. To the best of our knowledge, we are the first work to exploit superquadrics for both template-free object reconstruction and interaction recognition.
3. We extend two egocentric hand-object datasets by introducing new compositional splits and investigate compositional action recognition where a subset of action verb and noun combinations do not exist during training.
4. We achieve state-of-the-art performance on two public datasets, H2O and FPHA, in both official and compositional settings.

Related Work

Our work tackles the joint problem of 3D hand-object reconstruction and action recognition from egocentric RGB videos. We first review the literature on reconstructing and recognising *Hand-object interactions* from RGB inputs. Then, we provide a brief review on *Compositional action recognition* and *Superquadrics*.

Reconstructing hand-object interactions. While most existing research has primarily focused on single-hand (Boukhayma, Bem, and Torr 2019; Simon et al. 2017; Zimmermann and Brox 2017) or objects (Wang et al. 2019; Lepetit, Pilet, and Fua 2004; Zheng et al. 2022, 2024) in isolation, recently there has been a surge in interest of joint understanding of hand-object pose estimation. As the problem of reconstructing both hand and object is extremely ill-posed due to heavy mutual occlusions, many works (Tse et al. 2022b; Liu et al. 2021; Hampali et al. 2022; Yang et al. 2022) reduce this problem to 6D pose estimation with instance-specific templates. Meanwhile, some previous efforts (Hasson et al. 2019; Tse et al. 2022a; Chen et al. 2022; Ye et al. 2023a; Chen et al. 2023) do not assume to have access to ground-truth object models at test time and follow a template-free paradigm. However, these approaches would fail on unseen object instances as they either lack geometrical prior or overfit to a limited number of training objects. Our work is a template-free approach where we utilise primitive shape information from estimated superquadrics, and therefore it is more generalisable to unseen object instances for understanding hand-object interaction.

Recognising hand-object interactions. Action recognition is one of the most actively researched areas in computer vision (Jhuang et al. 2013; Varol, Laptev, and Schmid 2017; Kantorov and Laptev 2014) and significant progress has been made with the availability of large-scale datasets (Sigurdsson et al. 2016; Kay et al. 2017; Karpathy et al. 2014; Grauman et al. 2022). Here, we focus on methods (Tekin, Bogo, and Pollefeys 2019; Yang et al. 2020) that simultaneously estimate hand(-object) poses and interactions from egocentric videos. Garcia-Hernando et al. (2018) presents the first egocentric dataset and shows that 3D hand poses are beneficial for recognising actions. As the dataset contains visible magnetic sensors and does not include two-hand poses, a markerless dataset named H2O is developed to provide rich 3D annotations for egocentric 3D interaction recognition (Kwon et al. 2021). This enables recent research to develop unified models for understanding hand-object interactions. HTT (Wen et al. 2023) and H2OTR (Cho et al. 2023) are two closely related works that are based on a transformer architecture. HTT focuses on leveraging different temporal granularity information, while H2OTR exploits contact map for robust estimation. All of the above methods share a common approach by following the semantic relationship between hand poses and object labels to infer action. In contrast, we propose a collaborative learning framework which is designed explicitly to recognise interactions in a compositional fashion.

Compositional action recognition. This task is designed to alleviate the problem of appearance bias by disjointing the combination of actions and objects between training and testing. STIN (Materzynska et al. 2020) models actions as transformation of geometric relations in both spatial and temporal domains. This approach generalises well to most actions but fails when there are intrinsic state changes of objects. Kim and Hager (2020) proposes to fuse RGB information with instance bounding boxes to capture more complex actions. Sun et al. (2021) removes the appearance effect by counterfactual debiasing inference. While these methods have proven effective, they are primarily designed to leverage 2D geometric information and do not fully explore the potential of 3D geometric cues. As a result, their performance remains comparable to I3D (Carreira and Zisserman 2017). In contrast, we focus on leveraging 3D geometric cues for compositional action recognition.

Superquadrics recovery. Superquadric is a well-studied computational primitive shape abstraction, offering a diverse range of shape representations including cuboids, ellipsoids, cylinders, octohedra, and other variations. It was first proposed to model complex objects in computer graphics (Barr 1981). Solina and Bajcsy (1990) presents a method for abstracting simple objects from range images using a single superquadric. Subsequently, Leonardis, Jaklic, and Solina (1997); Chevalier, Jaillet, and Baskurt (2003) extend to recover more complex objects with multiple superquadrics. Recently, Liu et al. (2022) proposes a probabilistic approach to improve robustness to outlier and fitting accuracy. To the best of our knowledge, we are the first to leverage superquadrics for action recognition and demonstrate its effectiveness in in-the-wild scenarios.

Methodology

Our training pipeline, as shown in Fig. 2, takes a sequence of T RGB frames $\mathbf{I} \in \mathbb{R}^{T\times 256\times 256\times 3}$ of dynamic hands manipulating objects as input. We first obtain spatial features $\mathbf{x} \in \mathbb{R}^{T\times d}$ by passing each frame into a ResNet-18 (He et al. 2016) encoder where d refers to feature dimensions. To enhance the interaction between visual and geometric cues, we propose a simple collaborative learning framework by leveraging the Transformer encoder and decoder (Vaswani et al. 2017) as basic building blocks. Specifically, we design a two-branch network where the appearance branch extracts video features and the geometric branch aims to recover 3D hand-object geometric information. With such design, we can model appearance and geometric information at different temporal granularity. In addition, we explicitly model the compositionality of the interaction by decomposing the action class into a verb-and-noun pair. Finally, we combine video appearance and geometric representations for recognising egocentric hand-object interactions.

Appearance Branch

Given image features $\mathbf{x}$, we concatenate with a learnable token $\mathbf{T}_{\text{img}} \in \mathbb{R}^{d}$ and apply positional encoding before feeding them into a Transformer encoder. The encoder models the relationships of different spatial regions and $\mathbf{T}_{\text{img}}$ through self-attentions. The learnt token $\mathbf{T}^{*}_{\text{img}}$ captures essential global contexts from backbone representation which are used for compositional reasoning. In addition, the encoder outputs aggregated spatial representation $\mathbf{x}_{\text{appearance}} \in \mathbb{R}^{T\times d}$ which is later used for interaction recognition.

Geometric Branch

As estimating the poses of hand and object requires more local or nearby frames, we divide the video sequence into N consecutive segments. Specifically, we follow Wen et al. (2023) and use a shifting window strategy with window size t, *i.e.* $N = T/t$. The frames beyond sequence length T are padded but masked out from attention computation. Readers are referred to Wen et al. (2023) for more details.

Instead of aiming to reconstruct the hand and the manipulated object simultaneously, we first estimate the shape and poses of the manipulated object and leverage this geometric information to predict hand poses. The reason behind this approach is that joint estimation poses a significantly harder problem: First, self-occlusion and self-similarity between the joints of two hands are unique problems in interacting hands. Second, when interacting with objects, hands and objects often exhibit even greater occlusions. This problem is further amplified under egocentric view setting due to large degree of erratic camera motions.

In addition, existing work which relies on 2D/3D bounding boxes has limitations in accurately representing the

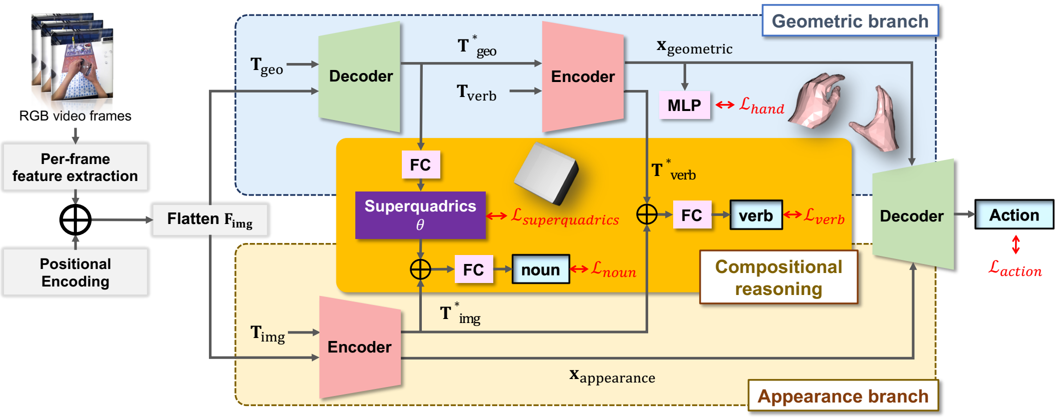

Figure 2: Overview of our approach. It first take RGB videos as input and produce per-frame spatial features $\mathbf{x}$ using a CNN backbone. Then, the appearance branch (bottom) applies positional encoding to $\mathbf{x}$ and combines it with a learnable token $\mathbf{T}_{\text{img}}$ before feeding into a Transformer encoder. Similarly, the geometric branch (top) extracts geometric features $\mathbf{T}_{\text{geo}}$ from flatten spatial features $\mathbf{F}_{\text{img}}$ using a Transformer decoder. The features from both branches are combined to predict superquadrics and object category. In addition, the geometric features are aggregated through another Transformer encoder to create global context-aware features between object shape and hand poses. The aggregated geometric features $\mathbf{x}_{\text{geometric}}$ and verb token features $\mathbf{T}_{\text{verb}}$ from this encoder are used to predict hand pose and action verb. Finally, the action class is predicted by feeding $\mathbf{x}_{\text{geometric}}$ into a cross-attention mechanism with the aggregated spatial representation $\mathbf{x}_{\text{appearance}}$ through a Transformer decoder.

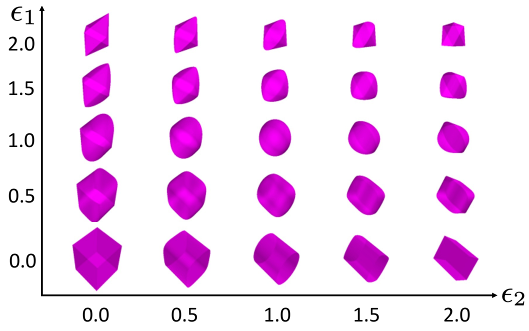

Figure 3: Qualitative examples of convex superquadrics. We show that superquadrics can model diverse objects by varying the shape parameters, ϵ_1 (y-axis) and ϵ_2 (x-axis).

shape and movement of objects. However, at the same time, it remains challenging to accurately reconstruct unseen objects from RGB images. Therefore, we propose to use superquadrics as a new object representation for improving action recognition. In the following, we present the *preliminaries* of superquadrics and detail our two-stage approach consisting *superquadrics decoder* and *hand pose estimator*.

Preliminaries. As shown in Fig. 3, superquadrics are a family of geometric primitives, *i.e.* cuboids, cylinders, ellipsoids, octahedra and their intermediates, which can be defined by an implicit function $f(\cdot)$ (Barr 1981):

$$f(\mathbf{p}) = \left(\left(\frac{x}{a_x}\right)^{\frac{2}{\epsilon_2}} + \left(\frac{y}{a_y}\right)^{\frac{2}{\epsilon_2}}\right)^{\frac{\epsilon_2}{\epsilon_1}} + \left(\frac{z}{a_z}\right)^{\frac{2}{\epsilon_1}} = 1, \tag{1}$$

where points $\mathbf{p} = [x, y, z] \in \mathbb{R}^3$ satisfying Eq. (1) form the surface of a superquadric. It can be encoded using 5 parameters: shape parameters $\epsilon_1, \epsilon_2 \in [0, 2] \subset \mathbb{R}$ and scale parameters $a_x, a_y, a_z \in \mathbb{R}_{>0}$. While the shape parameters can exceed 2 and result in non-convex shapes, we limit them within the convex region in this paper. We can now fully parameterise a superquadric by including the Euclidean transformation $g \in SE(3)$, *i.e.* $g = [\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3]$.

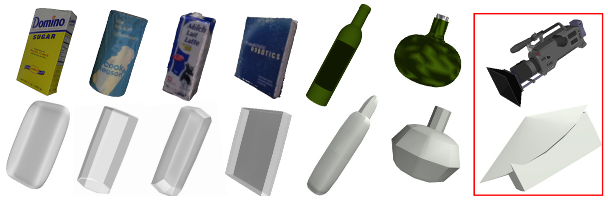

Figure 4: Qualitative examples of superquadrics. We extract superquadrics from everyday objects obtained from *YCB* (Calli et al. 2015), *ShapeNet* (Chang et al. 2015a), *FPHA* (Garcia-Hernando et al. 2018) and *H2O* (Kwon et al. 2021) datasets. We show that superquadrics have sufficient expressiveness to represent everyday objects. We also present an example failure case in the red box.

Superquadrics decoder. We use superquadrics $\boldsymbol{\theta}$ to be the intermediate 3D object representation, as it offers a compact way to represent a wide range of geometric primitives, *i.e.* $\boldsymbol{\theta} = \{\epsilon_1, \epsilon_2, a_x, a_y, a_z, g\} \in \mathbb{R}^{11}$. As shown in Fig. 4, superquadrics can provide sufficient expressiveness to reasonably model a diverse range of everyday objects. To this end, we extract geometric features $\mathbf{T}_{\text{geo}} \in \mathbb{R}^{N \times d}$ from the segmented spatial features using a Transformer decoder. We train a fully-connected layer to predict $\boldsymbol{\theta}$ from the learnt flatten $\mathbf{T}^*_{\text{geo}}$ by minimising the L1 loss $\mathcal{L}_{\text{superquadrics}}$. The main motivation to use Transformer decoder here instead of encoder is to leverage the autoregressive self-attention mechanism for modelling objects with multiple superquadrics. In addition, having separate encoder and decoder for this branch enables pre-training on large-scale object datasets that lack hand pose annotations.

Hand pose estimator. Given the learnt geometric features $\mathbf{T}^*_{\text{geo}}$, we concatenate with a learnable verb token $\mathbf{T}_{\text{verb}}$ and use self-attention in a Transformer encoder to create global context-aware features between object shape and hand poses. This encoder outputs aggregated geometric features $\mathbf{x}_{\text{geometric}} \in \mathbb{R}^{N \times d}$ and verb token features $\mathbf{T}^*_{\text{verb}} \in \mathbb{R}^d$.

We map the aggregated geometric features $\mathbf{x}_{\text{geometric}}$ to hand pose space by a 3-layer MLP and use MANO joint angles (Romero, Tzionas, and Black 2017) for hand pose representation. Specifically, we estimate 16 3D joint angles under the hand kinematic tree and MANO hand shape parameter per hand. Then, we can compute the 21 root-relative 3D joint locations of each hand by using the predicted joint angles and hand shape parameters. They are learned by minimising L1 loss $\mathcal{L}_{\text{hand}}$.

Compositional Reasoning

In the following, we describe how we leverage the compositional nature of actions by exploiting the action class as verb-noun pair with 3D geometric cues, *i.e.* superquadrics $\boldsymbol{\theta}$ and geometric-aware verb token $\mathbf{T}^*_{\text{verb}}$.

Object category predictor. We first predict the category of the manipulated object as it corresponds to the noun of an action. To achieve that, we leverage the basic primitive geometric information from superquadrics $\boldsymbol{\theta}$ as object shape provides strong signals to estimating object category. More specifically, we predict the classification probability vector for object category by linearly projecting the concatenation of superquadrics $\boldsymbol{\theta}$ and $\mathbf{T}^*_{\text{img}}$. We supervise this linear layer by minimising the cross-entropy loss $\mathcal{L}_{\text{noun}}$.

Verb predictor. Similarly, we predict action verb by feeding the concatenation of $\mathbf{T}^*_{\text{verb}}$ and $\mathbf{T}^*_{\text{img}}$ to a linear layer. It is also trained by minimising the cross-entropy loss $\mathcal{L}_{\text{verb}}$.

Discussion. The key idea for concatenating with $\mathbf{T}^*_{\text{img}}$ is to allow action verb and noun to interact with the appearance branch. Also, it generates loss gradients for both branches to develop a collaborative learning relationship. In addition, the motivation for estimating superquadrics first in the geometric branch is based on the fact that the human visual system flavours abstracting scenes into canonical parts for better perceptual understanding (Liu et al. 2022). This enables robust action recognition using basic geometric primitives instead of relying on accurate point-wise estimation. In summary, our design targets the problem of recognising a seen action when facing new objects by enabling the network to capture the compositionality of an action explicitly.

Interaction recognition

Besides explicitly modelling the compositionality of actions, our proposed framework can easily combine with any video-level appearance representation. The impact of appearance features can be two-fold: 1) The presence of appearance features can be particularly beneficial for action classes that lack prominent inter-object dynamics (Materzynska et al. 2020). 2) Conversely, appearance bias can inhibit the model learning ability by making strong correlations on spatial appearance rather than temporal or geometric transformations (Sun et al. 2021). To overcome the limitations of existing methods that can only accept or reject appearance information as a whole, we propose to use a Transformer decoder for recognising interaction. There are two key motivations for this design: 1) Transformer decoder has skip connections by design so it can attend segments of appearance features. 2) It can take different temporal granularity from both branches as inputs to make action predictions. Specifically, this decoder takes the aggregated geometric and spatial features, *i.e.* $\mathbf{x}_{\text{geometric}}$ and $\mathbf{x}_{\text{appearance}}$ as input and extracts relevant image features through cross-attention between geometric features. The vector output of this decoder is fed to a 3-layer MLP classifier of width and is supervised with cross-entropy loss $\mathcal{L}_{\text{action}}$. We investigate and analyse our design choices and provide details for the architecture in the supplementary materials.

Experiments

Implementation details. We train our model with the Adam optimiser (Kingma and Ba 2015) using an initial learning rate of 3×10^{-5} which halves in every 15 epochs. We keep the relative weights between different losses and normalise them for all experiments. We use ResNet (He et al. 2016) pre-trained on ImageNet (Russakovsky et al. 2015) for our backbone. For all Transformer encoders and decoders, we use 2 encoding/decoding layers where each layer has 8 attention heads. We use the fixed sine/cosine functions for positional encoding and add layer normalisation before the attention and feed-forward computations (Vaswani et al. 2017). We follow Wen et al. (2023) by setting $T = 128$, $t = 16$, $d = 512$ and training for 45 epochs with batch size of 2. Our final loss $\mathcal{L}_{\text{final}}$ is defined as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{superquadrics}} + \mathcal{L}_{\text{hand}} + \mathcal{L}_{\text{noun}} + \mathcal{L}_{\text{verb}} + \mathcal{L}_{\text{action}}. \quad (2)$$

Datasets. We conduct experiments on 3 interacting hand-object datasets and detail below.

ObMan (Hasson et al. 2019) is a synthetic dataset which was produced by rendering hand meshes with selected objects from ShapeNet (Chang et al. 2015b) dataset. It captures 8 object categories of everyday objects (*e.g.* bottles, cans and jars) and results in a total of 2,772 meshes. We precomputed superquadrics for all object meshes using the EMS algorithm (Liu et al. 2022) and pretrained the geometric branch on *ObMan* before training on other real datasets.

First-person hand benchmark (FPHA) (Garcia-Hernando et al. 2018) is a real dataset which records egocentric videos on diverse hand-object interactions. It captures 6 subjects performing 45 actions by interacting with 26 objects, *i.e.* juice bottle, liquid soap, milk and salt. We evaluate on the *action split* where all subjects and actions are present in both training and testing. This split consists of 600 and 575 videos for training and testing, respectively.

H2O (Kwon et al. 2021) is a recent real dataset which provides markerless 3D annotations for two hands and the

6D pose of manipulated objects. It is the first unified dataset for egocentric interaction recognition with rich 3D annotations of 4 subjects performing 36 actions on 8 objects. We follow Wen et al. (2023); Kwon et al. (2021) and use the sequences of egocentric view for training and testing.

Figure 5: Qualitative examples on *H2O*. We show that our model can recover plausible interaction across different object categories and hand-object configurations without object templates. We present additional qualitative examples and failure cases in the supplementary.

Baselines. We compare our method against MViTv2 (Li et al. 2022), HTT (Wen et al. 2023) and H2OTR (Cho et al. 2023). MViTv2 is widely adopted for video recognition tasks and can serve as a strong appearance-based baseline. We train the base variant of MViTv2 with weights pretrained on Kinetics-400 dataset (Kay et al. 2017) using the PySlowFast library (Fan et al. 2020). For the pose-based baselines, we consider HTT and H2OTR as they are recent methods that achieves state-of-the-art performance on both FPHA and H2O datasets. We also consider two Transformer-based baselines which do not contain compositional reasoning and superquadrics, respectively. We provide more details about the baselines in the supplementary materials. They are useful for understanding the importance of superquadrics and compositional reasoning for recognising interactions with unseen objects.

Evaluation metrics. We report the Mean End-Point Error (MEPE) in mm to evaluate *pose estimation*. MEPE measures the mean Euclidean distances between predictions and ground-truths. We also report the top-1 classification accuracy for *action recognition*.

Quantitative Comparison

Comparison with state-of-the-art. We report quantitative comparisons with the state-of-the-art methods on *H2O* and *FPHA* datasets in Table 1. We split the table into two sections: appearance-based (Wang et al. 2018; Carreira and Zisserman 2017; Feichtenhofer et al. 2019; Li et al. 2022) and geometric (Tekin, Bogo, and Pollefeys 2019; Yang et al. 2020; Kwon et al. 2021; Wen et al. 2023; Cho et al. 2023) methods for clear comparison. By considering the performances of action recognition, the appearance-based baseline MViTv2 performs competitively with the state-of-the-art geometric method H2OTR (Cho et al. 2023). It raises an immediate question as to when 3D geometric cues be beneficial to recognising interaction as collecting ground-truth contact maps or other 3D annotations are non-trivial (Brahmbhatt et al. 2019; Tse et al. 2022b; Kwon et al. 2021). We will address this question in the following paragraph. Nonetheless, we demonstrate the effectiveness of explicit compositional reasoning with superquadrics by outperforming all methods. On the other hand, for pose estimations, H2OTR is a Transformer-based framework which achieves state-of-the-art accuracy for hand pose estimation. However, all of the compared methods cannot recover dense object geometries without manual selection of object models at test time. In contrast, our method performs competitively without known object templates and outperforms all methods on object pose estimation. We attribute this to the fact that superquadrics can provide dense 3D geometric information about the manipulated object, whereas 2D or 3D bounding boxes have limitations on representing object shape and movement. We show qualitative results on *H2O* dataset in Fig. 5.

Compositional action recognition. We further evaluate on the compositional recognition task in Table 2. Following Materzynska et al. (2020), we first create new splits for the task of compositional action recognition by extending existing egocentric hand-object datasets. Specifically, we remove the sequences that contain the predefined object category from the train split, such that the combinations of a verb (action) and nouns do not overlap in the testing set. To gain a deeper understanding of the model generalisation ability on unseen objects, we evaluate the model using $N_{\text{obj}}-$fold cross validation where N_{obj} refers to the number of total objects presented in the dataset. We keep the original testing splits (named $\mathcal{S}_0$) to illustrate the difficulty of this compositional task. We further experiment on a more challenging split where two object categories are randomly removed in the *H2O* dataset. We name the base splits by $\mathcal{S}_1$ and the more difficult splits where additional verb-nouns combinations are removed from training by $\mathcal{S}_2$. We report the mean and the standard deviation of top-1 classification accuracy for all experiments. We present the main results in Table 2, while full results are presented in the supplementary.

As shown in Table 2, our proposed collaborative learning framework consistently outperforms both the appearance and geometric baselines, *i.e.* MViTv2, HTT and H2OTR. We find that the performance of MViTv2 drastically drops by 29.44% and 37.76%, in $\mathcal{S}_1$ and $\mathcal{S}_2$ of H2O respectively. These results are in line with previous studies (Materzynska et al. 2020; Zhou et al. 2018; Sun et al. 2021) where deep architectures tend to overfit the object appearance. By adding geometric cues in HTT and H2OTR, we observe a small performance gain by an average of 7.75% and 9.95% on both compositional splits, respectively. We further evaluate hand pose estimation accuracy under this compositional setting. Similarly, we report the mean and the standard deviation of MEPE in mm for all experiments. By comparing with HTT, we achieve state-of-the-art performance in hand pose estimation with the advantage of object reconstruction

Method	*H2O* Top-1 accuracy (%) ↑	*H2O* Hand (L/R) ↓	*H2O* Obj. ↓	*FPHA* Top-1 accuracy (%) ↑	*FPHA* Hand ↓	*FPHA* Obj. ↓
C2D (Wang et al. 2018)	70.66	-	-	-	-	-
I3D (Carreira and Zisserman 2017)	75.21	-	-	-	-	-
SlowFast (Feichtenhofer et al. 2019)	77.69	-	-	-	-	-
MViTv2 (Li et al. 2022)	90.08	-	-	98.45	-	-
H+O (Tekin, Bogo, and Pollefeys 2019)	68.88	41.4/38.9	50.4	82.43	15.8	24.9
H2O (Kwon et al. 2021)	79.25	41.5/37.2	47.9	-	-	-
HTT (Wen et al. 2023)	86.36	35.0/36.1	-	94.09	-	-
H2OTR (Cho et al. 2023)	90.90	**24.4/25.8**	45.2	98.4	15.0	21.0
Ours	**92.25**	28.9/30.2	**43.5**	**98.74**	**13.6**	**20.1**

Table 1: Error rates of action recognition and pose estimation on *H2O* and *FPHA*. We report the top-1 accuracy for classification and MEPE in mm for hand (left/right) and object error. Our proposed method performs competitively without known object templates at inference.

	H2O						*FPHA*			
Method	$\mathcal{S}_0(\%)$ ↑	Hand ↓	$\mathcal{S}_1(\%)$ ↑	Hand ↓	$\mathcal{S}_2(\%)$ ↑	Hand ↓	$\mathcal{S}_0(\%)$ ↑	Hand ↓	$\mathcal{S}_1(\%)$ ↑	Hand ↓
MViTv2 (Li et al. 2022)	90.08	-	60.64±2.3	-	52.32±5.7	-	98.45	-	69.02±1.8	-
HTT (Wen et al. 2023)	86.36	35.6	71.13±2.7	37.4±2.3	59.88±2.6	41.5±1.7	94.09	15.8	74.21±1.8	18.8±1.2
H2OTR (Cho et al. 2023)	90.90	25.1	73.54±67	35.6±3.7	61.78±6.1	40.2±5.2	98.4	15	76.52±9.1	18.3±4.2
Ours	**92.25**	29.6	**80.59±1.6**	**31.8±2.1**	**69.93±2.5**	**33.4±1.5**	**98.74**	**13.6**	**85.80±1.4**	**13.9±0.9**

Table 2: Error rates of compositional action recognition *H2O* and *FPHA*. We report classification accuracy in $\%$ and hand error in mm for the official split $\mathcal{S}_0$ and two additional compositional split settings, $\mathcal{S}_1$ and $\mathcal{S}_2$. The results are reported as the mean and standard deviation of the performance metric of interest, providing a comprehensive understanding of the model's performance across different split settings.

Method	Verb(%) ↑	Noun(%) ↑	Verb+Noun(%) ↑	Top-1(%) ↑	Hand(mm) ↓
w/o geometric branch	-	-	-	78.91	-
w/o compositional reasoning	-	-	-	81.82	38.91
w/o verb classifier	-	85.61	-	83.45	37.15
w/o noun classifier	86.24	-	-	83.96	37.62
w/o interaction decoder	88.07	90.04	87.2	-	32.67
w/o superquadrics	90.18	91.56	90.01	89.85	31.85
Ours	**92.23**	**96.89**	**92.25**	**90.0**	**29.6**

Table 3: Ablation study of model architecture design on *H2O*. We report top-1 classification accuracy $(\%)$ and MEPE for hand error in mm. In addition, we include classification predictions for verb and noun from compositional reasoning.

using superquadrics. Our strong performances across all settings demonstrate the importance of explicit reasoning about interactions with 3D geometric information.

Ablations

To motivate our design choices, we perform additional qualitative evaluation of our method with various key components disabled in Table 3. We evaluate the effectiveness of our collaborative learning framework by experimenting on a single appearance branch (row 1) and a two-branch network without gradient flow (row 2), which yields the lowest performance in *H2O*. Then, we observe significant performance drops by removing either one of the inter-branch classifiers (row 3, 4). These results demonstrate the effectiveness of our collaborative learning framework which encourages information sharing between two-branches. Further, we are interested in finding out whether incorporating the interaction decoder is important as action class can be obtained by combining verb and noun predictions. We show that the interaction decoder can bring performance gain for verb and noun classifiers by additional supervision in row 5. Finally, we show that superquadric predictions (*i.e.* by removing the purple block in Fig. 2) can push the limit to the new state-of-the-art in all metrics.

Conclusion

We showed that, by explicitly leveraging 3D geometric information, we could recognise actions performed on unseen objects much more accurately than existing state-of-the-arts. We also demonstrated that superquadrics as a new object representation for action recognition to be effective. We validated our approach by extending existing datasets with compositional splits and achieved state-of-the-art performance. We plan to investigate more complex interactions with articulated objects in the future.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support programme (IITP-2024-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work is supported in part by the National Natural Science Foundation of China under Grant No.62203184.

References

Barr, A. H. 1981. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261*.

Boukhayma, A.; Bem, R. d.; and Torr, P. H. 2019. 3D hand shape and pose from images in the wild. In *CVPR*.

Brahmbhatt, S.; Ham, C.; Kemp, C. C.; and Hays, J. 2019. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*.

Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. M. 2015. The YCB object and model set: Towards common benchmarks for manipulation research. In *ICAR*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015a. ShapeNet: An Information-Rich 3D Model Repository. (arXiv:1512.03012).

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015b. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*.

Chen, Z.; Chen, S.; Schmid, C.; and Laptev, I. 2023. gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction. In *CVPR*.

Chen, Z.; Hasson, Y.; Schmid, C.; and Laptev, I. 2022. AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*.

Chevalier, L.; Jaillet, F.; and Baskurt, A. 2003. Segmentation and superquadric modeling of 3D objects.

Cho, H.; Kim, C.; Kim, J.; Lee, S.; Ismayilzada, E.; and Baek, S. 2023. Transformer-Based Unified Recognition of Two Hands Manipulating Objects. In *CVPR*.

Fan, H.; Li, Y.; Xiong, B.; Lo, W.-Y.; and Feichtenhofer, C. 2020. PySlowFast. https://github.com/facebookresearch/slowfast.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *ICCV*.

Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *CVPR*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast networks for video recognition. In *ICCV*.

Feng, R.; Gao, Y.; Ma, X.; Tse, T. H. E.; and Chang, H. J. 2023a. Mutual Information-based Temporal Difference Learning for Human Pose Estimation in Video. In *CVPR*.

Feng, R.; Gao, Y.; Tse, T. H. E.; Ma, X.; and Chang, H. J. 2023b. DiffPose: SpatioTemporal Diffusion Model for Video-based Human Pose Estimation. In *ICCV*.

Garcia-Hernando, G.; Yuan, S.; Baek, S.; and Kim, T.-K. 2018. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *ICCV*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Hampali, S.; Sarkar, S. D.; Rad, M.; and Lepetit, V. 2022. Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. In *CVPR*.

Han, S.; Liu, B.; Cabezas, R.; Twigg, C. D.; Zhang, P.; Petkau, J.; Yu, T.-H.; Tai, C.-J.; Akbay, M.; Wang, Z.; et al. 2020. MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ToG*.

Han, S.; Wu, P.-c.; Zhang, Y.; Liu, B.; Zhang, L.; Wang, Z.; Si, W.; Zhang, P.; Cai, Y.; Hodan, T.; et al. 2022. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH*.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *ICCV*.

Kantorov, V.; and Laptev, I. 2014. Efficient feature extraction, encoding and classification for action recognition. In *CVPR*.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv:1705.06950*.

Kim, T. S.; and Hager, G. D. 2020. SAFCAR: Structured attention fusion for compositional action recognition. *arXiv:2012.02109*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kwon, T.; Tekin, B.; Stühmer, J.; Bogo, F.; and Pollefeys, M. 2021. H2O: Two hands manipulating objects for first person interaction recognition. In *ICCV*.

Leonardis, A.; Jaklic, A.; and Solina, F. 1997. Superquadrics for segmenting and modeling range data. *TPAMI*.

Lepetit, V.; Pilet, J.; and Fua, P. 2004. Point matching as a classification problem for fast and robust object pose estimation. In *CVPR*.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. MViVv2: Improved multiscale vision transformers for classification and detection. In *CVPR*.

Liu, S.; Jiang, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*.

Liu, W.; Wu, Y.; Ruan, S.; and Chirikjian, G. S. 2022. Robust and accurate superquadric recovery: A probabilistic approach. In *CVPR*.

Materzynska, J.; Xiao, T.; Herzig, R.; Xu, H.; Wang, X.; and Darrell, T. 2020. Something-Else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6): 245.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. ImageNet large scale visual recognition challenge. *IJCV*.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *NeurIPS*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *NeurIPS*.

Solina, F.; and Bajcsy, R. 1990. Recovery of parametric models from range images: The case for superquadrics with global deformations.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*.

Sun, P.; Wu, B.; Li, X.; Li, W.; Duan, L.; and Gan, C. 2021. Counterfactual debiasing inference for compositional action recognition. In *ACM MM*.

Tekin, B.; Bogo, F.; and Pollefeys, M. 2019. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *CVPR*.

Tse, T. H. E.; Kim, K. I.; Leonardis, A.; and Chang, H. J. 2022a. Collaborative Learning for Hand and Object Reconstruction with Attention-guided Graph Convolution. In *CVPR*.

Tse, T. H. E.; Mueller, F.; Shen, Z.; Tang, D.; Beeler, T.; Dou, M.; Zhang, Y.; Petrovic, S.; Chang, H. J.; Taylor, J.; and Doosti, B. 2023. Spectral Graphormer: Spectral Graph-Based Transformer for Egocentric Two-Hand Reconstruction using Multi-View Color Images. In *ICCV*.

Tse, T. H. E.; Zhang, Z.; Kim, K. I.; Leonardis, A.; Zheng, F.; and Chang, H. J. 2022b. S^2Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning. In *ECCV*.

Varol, G.; Laptev, I.; and Schmid, C. 2017. Long-term temporal convolutions for action recognition. *TPAMI*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; and Savarese, S. 2019. DenseFusion: 6D object pose estimation by iterative dense fusion. In *CVPR*.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*.

Wen, Y.; Pan, H.; Yang, L.; Pan, J.; Komura, T.; and Wang, W. 2023. Hierarchical temporal transformer for 3D hand pose estimation and action recognition from egocentric RGB videos. In *CVPR*.

Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022. ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis. In *CVPR*.

Yang, S.; Liu, J.; Lu, S.; Er, M. H.; and Kot, A. C. 2020. Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis. In *ECCV*.

Ye, Y.; Hebbar, P.; Gupta, A.; and Tulsiani, S. 2023a. Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips. In *ICCV*.

Ye, Y.; Li, X.; Gupta, A.; De Mello, S.; Birchfield, S.; Song, J.; Tulsiani, S.; and Liu, S. 2023b. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*.

Zheng, L.; Leonardis, A.; Tse, T. H. E.; Horanyi, N.; Chen, H.; Zhang, W.; and Chang, H. J. 2022. TP-AE: Temporally Primed 6D Object Pose Tracking with Auto-Encoders. In *ICRA*.

Zheng, L.; Tse, T. H. E.; Wang, C.; Sun, Y.; Chen, H.; Leonardis, A.; Zhang, W.; and Chang, H. J. 2024. GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pose Refinement. In *CVPR*.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal relational reasoning in videos. In *ECCV*.

Zimmermann, C.; and Brox, T. 2017. Learning to estimate 3d hand pose from single rgb images. In *ICCV*.