EMAG: Ego-motion Aware and Generalizable 2D Hand Forecasting from Egocentric Videos

Masashi Hatano[1], Ryo Hachiuma[2], and Hideo Saito[1]

[1] Keio University
[2] NVIDIA

Abstract. Predicting future human behavior from egocentric videos is a challenging but critical task for human intention understanding. Existing methods for forecasting 2D hand positions rely on visual representations and mainly focus on hand-object interactions. In this paper, we investigate the hand forecasting task and tackle two significant issues that persist in the existing methods: (1) 2D hand positions in future frames are severely affected by ego-motions in egocentric videos; (2) prediction based on visual information tends to overfit to background or scene textures, posing a challenge for generalization on novel scenes or human behaviors. To solve the aforementioned problems, we propose EMAG, an ego-motion-aware and generalizable 2D hand forecasting method. In response to the first problem, we propose a method that considers ego-motion, represented by a sequence of homography matrices of two consecutive frames. We further leverage modalities such as optical flow, trajectories of hands and interacting objects, and ego-motions, thereby alleviating the second issue. Extensive experiments on two large-scale egocentric video datasets, Ego4D and EPIC-Kitchens 55, verify the effectiveness of the proposed method. In particular, our model outperforms prior methods by 1.7% and 7.0% on intra and cross-dataset evaluations, respectively. Project Page: `https://masashi-hatano.github.io/EMAG/`
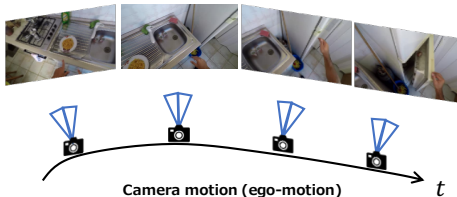
Keywords: Egocentric Vision · 2D Hand Forecasting

1 Introduction

With the emergence of wearable devices such as smart glasses and intelligent helmets, there has been growing interest in the analysis of egocentric videos. In recent years, large-scale egocentric vision datasets such as EPIC-Kichtens [8, 9] and Ego4D [16] have been introduced to catalyze the next era of research in first-person perception and provide a diverse range of tasks for investigation, including action recognition [15, 36, 53], human body pose estimation [26, 51, 52], audio-visual understanding [21, 41], action anticipation [14, 38], and natural language queries [40].

Future forecasting is one of the major categories, including the anticipation of the camera wearer's future actions and the prediction of human movements. This capability has immediate applications in AR/VR [56, 57] and human-robot

interactions [39,55] as both fields benefit from understanding the camera wearer's actions or behaviors. Among the tasks in future forecasting, hand forecasting has been recognized as particularly challenging due to severe ego-motion, which affects the 2D hand positions in future frames.

Recent 2D egocentric hand forecasting approaches [16, 30, 31] leverage visual feature representations extracted from input RGB videos using 2D or 3D Convolutional Neural Networks (CNNs) for the hand forecasting task. For example, the method proposed in the Ego4D dataset [16] uses a simple I3D network [5] and regresses the future 2D hand coordinates. Meanwhile, the Object Centric Transformer (OCT) [31] is a method that jointly predicts hand motions

Fig. 1: The presence of ego-motion in first-person videos significantly affects the dynamic movement of the camera position. Since the camera is part of the wearer's body, a variety of views can be captured even in a short period of time.

and object contact points from RGB video features extracted with BNInception [48] and the hand/object bounding boxes.

Although the 2D hand forecasting task has been widely studied, two critical issues still remain in the previous works: the accuracy and generalization performance against unseen data, both of which are crucial for practical scenarios. First, the 2D hand position in future frames is heavily influenced by the head motion of the camera wearer, also known as *ego-motion*. As illustrated in Fig. 1, body and head motions cause frequent view changes even in a short period of time, yet the previous approaches have not explicitly considered ego-motion for predicting 2D hand positions. Second, the performance of RGB-based prediction approaches significantly drops when the video feature distribution (*i.e.* domain) diverges from that of the training set [25, 54]. This performance drop is crucial for the 2D egocentric hand forecasting task since the camera is not situated at a fixed location. For instance, performance may vary if the egocentric videos are captured in different textured environments (*e.g.*, outdoor vs. indoor), or if the wearer performs different actions from the training.

This work proposes *EMAG*, an ego-motion-aware and generalizable 2D hand forecasting method. This approach capitalizes on the incorporation of ego-motion information to enhance the accuracy of the hand forecasting task. Additionally, we employ multiple modalities to mitigate susceptibility to overfitting in backgrounds or scene textures. We aim to achieve more robust predictions in settings where camera wearers engage in a diverse range of tasks such as cooking and gardening.

To address the first challenge, we propose leveraging a sequence of homography matrices as ego-motion and anticipating them on future frames. Given that hand positions in future frames are affected by future ego-motion, explicitly forecasting ego-motion as an auxiliary task enhances the accuracy of predict-

ing future hand positions, particularly in egocentric videos where head motions occur frequently.

To alleviate the second issue, instead of primarily relying on visual features for estimating 2D hand positions, we leverage modalities such as optical flow, hand/object positions, and ego-motion information, using hand motions as the primary features for hand forecasting. This approach reduces reliance on appearance-based features, as these modalities are free from appearance-based biases [36]. Consequently, the model's generalizability is enhanced, ensuring robust performance even when distribution gaps exist between the training and test data.

We extensively evaluate the proposed method on two large-scale egocentric datasets, Ego4D [16] and EPIC-Kitchens 55 [8]. The performance of the proposed method, along with that of previous state-of-the-art forecasting approaches, is assessed under two settings: the intra-dataset setting and the cross-dataset setting. In the cross-dataset setting, the model is evaluated on a different dataset from training to verify the generalization performance against unseen scenes or actions. As a result, our method outperforms the previous approaches in both two settings (1.7% and 7.0% improvement with intra-dataset and cross-dataset settings, respectively). Moreover, we conduct various ablation studies on the proposed input modalities and loss components.

In summary, our contributions are as follows:

- We are the first to investigate the potential benefits of incorporating ego-motion, which is critical in the 2D hand forecasting task.
- We propose a simple but effective approach, EMAG, that considers ego-motion, represented by a sequence of homography matrices of two consecutive frames. In addition, our method utilizes multiple modalities to mitigate overfitting to scene textures.
- We conduct extensive experiments on two large-scale egocentric datasets, Ego4D and EPIC-Kitchens 55. The experimental results verify the outperformance of the proposed method over the previous approaches through two different experimental setups: intra-dataset and cross-dataset. Especially, the method shows strong performance with cross-dataset in which the training and test datasets differ.

2   Related Work

2.1   Egocentric Video Understanding

Video understanding is one of the central tasks in the computer vision field. Various video understanding methods are well-established thanks to large-scale datasets [17, 24, 45] collected from internet sources (*e.g.* YouTube). The videos in these large-scale datasets are mostly captured from an exocentric camera (third-view video), such as a surveillance or a hand-held camera.

On the other hand, analyzing egocentric video (first-view video) captured by wearable cameras has become an active area of research in recent years [7, 27, 29,

35, 37, 58]. Compared with exocentric videos, egocentric videos provide distinct viewpoints of surrounding scenes and actions driven by the camera position holding on the observer. Therefore, egocentric video analysis can be helpful for various applications, such as AR/VR [56, 57] or medical image analysis [4, 12].

Multiple large-scale egocentric video datasets [8, 9, 16, 28, 32] have been proposed in response to the demand for egocentric video analysis. These datasets have played a pivotal role in advancing research on egocentric video understanding, encompassing tasks such as activity recognition [15, 36, 53], human-object interaction [30, 31, 59], action anticipation [14, 38], human body pose estimation [26, 51, 52], and audio-visual understanding [21, 41]. In this work, we explore one of the challenging tasks in egocentric video analysis, 2D hand forecasting.

### 2.2   Hand Forecasting from Egocentric Videos

To predict future hand positions, traditional tracking or sequential methods, such as Kalman Filter (KF) [23], Constant Velocity Model (CVM) [43], and Seq2Seq [47], have been commonly employed for trajectory prediction. These methods often rely solely on trajectories of hand positions and do not effectively leverage the context of scenes without visual information, resulting in suboptimal performance. To effectively leverage visual information, the baseline method for hand forecasting, which was proposed as a benchmark along with the Ego4D [16] dataset, utilized I3D [5], a method that is known for its outstanding performance to extract spatial and temporal information.

Moreover, several studies have focused on hand-object interactions to explore the relationship between meaningful human body movements and future representations. FHOI [30] is the first work to incorporate the future trajectory of hands for action anticipation in egocentric videos. Building upon this, OCT [31] is an approach that integrates hand-object interactions into the prediction process.

However, neither of these approaches explicitly considers ego-motion, which plays a crucial role in accurately predicting future hand positions in 2D image coordinates, as future hand positions are heavily influenced by future ego-motion. In contrast to previous works, we explore the potential benefits of integrating ego-motion information to enhance the capability of predicting future hand positions even in the presence of severe ego-motion.

3   Method

The proposed architecture is built upon the original Transformer [50]. It inputs multiple modalities and predicts future hand positions and ego-motions. We first introduce the egocentric 2D hand forecasting task (Sec. 3.1). Then, we introduce our proposed method, including pre-processing (Sec. 3.2), an encoder (Sec. 3.3), our hand position and ego-motion predictors (Sec. 3.4), and our training objective (Sec. 3.5). Fig. 2 provides an overview of our approach.
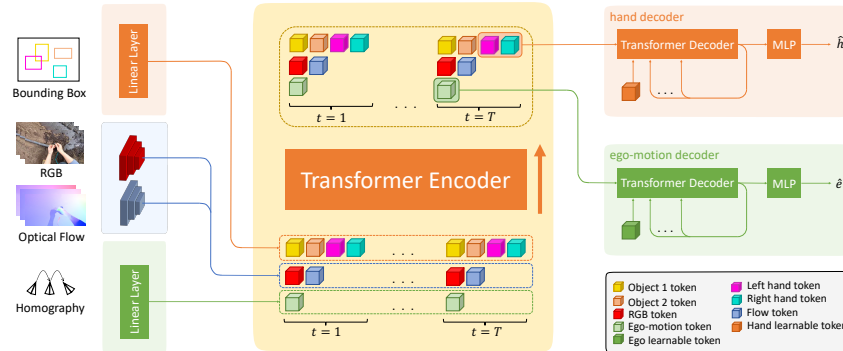
Fig. 2: The architecture of the proposed method. Given input egocentric video frames, we pre-process them and obtain multiple modalities, including RGB and optical flow, detected bounding boxes of objects/hands, and homography matrices of adjacent frames. We train a single Transformer encoder and two Transformer decoders with MLP heads for hand and ego-motion prediction.

3.1   Problem Definition

The task is to predict future hand positions of the camera wearer in 2D image coordinates on future frames, followed by the definition on Ego4D [16]. Given an input egocentric video $V = \{I^1, \ldots, I^T\}$ with an observation time length T, where I^T represents the last observation frame. Our goal is to predict future hand coordinates $\boldsymbol{h} = \{\boldsymbol{h}^{T+1}, \ldots, \boldsymbol{h}^{T+F}\}$ for the future time horizons F. At each time step t, $\boldsymbol{h}^t$ consists of left/right-hand positions in the 2D image coordinate system on the frame I^t.

3.2   Pre-processing

Our proposed method inputs three types of input modalities: trajectory information, global information, and ego-motion information. We pre-process an input video to obtain these three modalities as follows.

Trajectory information. Trajectory information consists of the sequential 2D positions of the bounding boxes of hands and objects. To obtain bounding boxes for both hands and objects for each frame, we apply an egocentric hand-object detector [44], which detects the left and right hand and objects separately. We use the following bounding boxes: left hand, right hand, and objects detected with a top-k confidence score.

Global information. Global information consists of RGB frames and optical flow. The optical flow can be estimated from two consecutive RGB frames via an off-the-shelf optical flow estimator, such as RAFT [49] or FlowFormer [22].

Ego-motion information. The ego-motion is represented by a sequence of homography matrices, which encapsulate the transformation between consecutive frames. Generally, a homography between images taken from two distinct viewpoints depends on the intricate 3D arrangement of the captured scene. Nonetheless, given the relatively small magnitude of the translation vector connecting consecutive frames in the context of first-person videos, a homography does not depend on the 3D structure of the scene but solely on the rotation between the two viewpoints.

The process of estimating the homography matrix involves two key stages: the identification of matching points between frames and the determination of a homography matrix that minimizes the error. The initial step entails identifying matching points, a task facilitated by using previously estimated optical flow, which characterizes the pixel displacement between frames. For the second step, we apply the RANSAC algorithm [11], which is known as a robust iterative algorithm, to estimate the homography parameters.

3.3   Encoder

Tokenization. After pre-processing all input modalities, each modality is transformed into a token to be encoded in a single Transformer encoder. For each detected bounding box (top-left and bottom-right coordinates) at time step t, it is transformed into a token $\boldsymbol{x}_i^t$ by a shared linear layer, which maps $\mathbb{R}^4 \to \mathbb{R}^C$, where i represents either of the left hand, right hand, or objects detected with a top-k confidence score, and C denotes the dimension size of each token. As for the global information, we use two 2D CNNs to extract the features of each RGB and flow frame and then pool the extracted features in the spatial direction by global average pooling (GAP). The pooled features are denoted as $\boldsymbol{x}_{\mathrm{rgb}}^t$ and $\boldsymbol{x}_{\mathrm{flow}}^t$. Similar to the trajectory information, each homography matrix is transformed into a token $\boldsymbol{x}_{ego}^t$ by a linear layer, which maps $\mathbb{R}^9 \to \mathbb{R}^C$. The 3×3 homography matrix is flattened before passing through the linear layer.

Index encoding. As there are various tokens in terms of modality type and time, two index encodings, the modal index embedding and time index embedding, are employed. The learnable position embedding is employed for the modal index embedding. Also, we adopt the time index encoding, which replaces the position in the original sinusoid positional encoding [50] with a time index (frame number).

Transformer encoder. We use a single Transformer encoder $\mathcal{E}$ to encode multiple input modalities across multiple time steps via self-attention mechanisms:

$$\boldsymbol{z}_{m_1}^1, \boldsymbol{z}_{m_2}^1, \ldots, \boldsymbol{z}_{m_M}^T = \mathcal{E}(\boldsymbol{x}_{m_1}^1, \boldsymbol{x}_{m_2}^1, \ldots, \boldsymbol{x}_{m_M}^T), \tag{1}$$

where $\boldsymbol{x}_{m_j}^t$ is the token of the m_j-th modality at the time step t, M denotes the number of input modality types, and $\boldsymbol{z}_{m_j}^t$ is the output token from the Transformer encoder $\mathcal{E}$.

3.4   Hand Position and Ego-motion Predictors

We use two Transformer decoders, the hand decoder ($\mathcal{D}_{\mathrm{hand}}$) and the ego-motion decoder ($\mathcal{D}_{\mathrm{ego}}$), conditioned on the features from the encoder in an autoregressive manner. Finally, the decoded feature for each future time step is fed into two MLP heads, $\mathcal{M}_{\mathrm{hand}}$ and $\mathcal{M}_{\mathrm{ego}}$, to predict the hand position and ego-motion for each time step.

Transformer decoder. For the hand Transformer decoder, the encoded left-hand token and the right-hand token of the last observation time T, $\boldsymbol{z}_{left}^{T}$ and $\boldsymbol{z}_{right}^{T}$, are used as the key and the value, and a learnable parameter is used as a hand learnable token $\boldsymbol{p}_{\mathrm{hand}}$ for the query of the first forecasting time step ($\boldsymbol{q}_{\mathrm{hand}}^{T} = \boldsymbol{p}_{\mathrm{hand}}$):

$$\boldsymbol{q}_{\mathrm{hand}}^{T+f} = \mathcal{D}_{\mathrm{hand}}(\boldsymbol{q}_{\mathrm{hand}}^{T}, \ldots, \boldsymbol{q}_{\mathrm{hand}}^{T+f-1}), \tag{2}$$

where $\boldsymbol{q}_{\mathrm{hand}}^{T+f}$ represents the decoded tokens for the future time step $T + f, f = \{1, \ldots, F\}$. We perform the same operation for the ego-motion Transformer decoder. The difference is the key, value, and query. The key and value stem from the encoded ego-motion features at the last observed time step T, $\boldsymbol{z}_{ego}^{T}$, and the query is a learnable parameter for ego-motion $\boldsymbol{p}_{\mathrm{ego}}$ ($= \boldsymbol{q}_{\mathrm{ego}}^{T}$):

$$\boldsymbol{q}_{ego}^{T+f} = \mathcal{D}_{\mathrm{ego}}(\boldsymbol{q}_{\mathrm{ego}}^{T}, \ldots, \boldsymbol{q}_{\mathrm{ego}}^{T+f-1}). \tag{3}$$

MLP head. We use multi-layer perceptrons (MLP), which take the decoded features from the Transformer decoder at each future time step for both hand position and ego-motion prediction. $\mathcal{M}_{\mathrm{hand}}$ predicts the coordinates of the left and right hands $\hat{\boldsymbol{h}}^{T+f}$ at the future time step $T + f$. Similarly, $\mathcal{M}_{\mathrm{ego}}$ predicts the nine elements of the homography matrix $\hat{\boldsymbol{e}}^{T+f}$:

$$\hat{\boldsymbol{h}}^{T+f} = \mathcal{M}_{\mathrm{hand}}(\boldsymbol{q}_{\mathrm{hand}}^{T+f}), \tag{4}$$

$$\hat{\boldsymbol{e}}^{T+f} = \mathcal{M}_{\mathrm{ego}}(\boldsymbol{q}_{\mathrm{ego}}^{T+f}). \tag{5}$$

Note that the weights of each MLP head ($\mathcal{M}_{\mathrm{hand}}$ and $\mathcal{M}_{\mathrm{ego}}$) are shared for each time step.

3.5   Training Objective

In our training process, we use two types of losses: the hand forecasting loss $\mathcal{L}_{\mathrm{hand}}$ and the ego-motion (nine elements of the homography matrix) estimation loss $\mathcal{L}_{\mathrm{ego}}$.

Hand forecasting loss. We adopt the self-adjusting smooth L1 loss, which was introduced in RetinaMask [13], as the objective function for hand forecasting:

$$l_i = \begin{cases} 0.5 w_i (h_i - \hat{h}_i)^2 / \beta, & |h_i - \hat{h}_i| < \beta \\ w_i(|h_i - \hat{h}_i| - 0.5\beta), & \text{otherwise} \end{cases} \tag{6}$$

$$\mathcal{L}_{\text{hand}} = \frac{1}{4F} \sum_i l_i, \tag{7}$$

where h_i is a i-th element of a vector representing the x, y ground truth coordinates of the left/right hands on F future frames $\boldsymbol{h} \in \mathbb{R}^{4F}$, $\hat{\boldsymbol{h}}$ denotes predicted future hand coordinates, and β is a control point that mitigates over-penalizing outliers. If the hand is not observed in future frames, we pad 0 into the $\hat{\boldsymbol{h}}$ and adopt a binary mask $\boldsymbol{w} \in \mathbb{R}^{4F}$ to prevent gradient propagation for these unobserved instances.

Ego-motion estimation loss. We employ the L2 loss for ego-motion estimation loss:

$$\mathcal{L}_{\text{ego}} = \frac{1}{9F} \sum_i (e_i - \hat{e}_i)^2, \tag{8}$$

where $\boldsymbol{e} \in \mathbb{R}^{9F}$ is a vector representing the elements of homography matrices on F future frames. $\mathcal{L}_{\text{hand}}$ and $\mathcal{L}_{\text{ego}}$ are linearly combined with a balancing hyperparameter α for the final training loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hand}} + \alpha \mathcal{L}_{\text{ego}}. \tag{9}$$

4   Experiments

4.1   Datasets

EPIC-Kitchens 55 [8]. EPIC-Kitchens 55 is the dataset that only includes the daily activities videos in the kitchen. It comprises a set of 432 egocentric videos recorded by 32 participants in their kitchens using a head-mounted camera. We use the train/val split provided by RULSTM [14].

Ego4D [16]. The Ego4D dataset is the most recent large-scale egocentric video dataset. It contains 3,670 hours of egocentric videos of people performing diverse tasks, such as farming or cooking, and is collected by 931 people from 74 locations across nine different countries worldwide. We follow the same train/val split protocol provided by Ego4D [16].

Followed by the previous work [31], we employ the egocentric hand-object detector [44] with the same setup as the previous work and consider the center of detected hand bounding boxes as the ground truth hand positions for both left and right hands.

4.2   Implementation Details

Experimental setup. We sample $T = 8$ frames at 4 FPS (frames per second) as input observations and forecast 1 second with the future time step $F = 4$ on both EPIC-Kitchens 55 and Ego4D. We use the pre-trained ResNet-18 [19] on ImageNet [10] as the backbone to extract RGB and optical flow features. We adopt the hand and object detector from the egocentric video [44] to detect left/right hand and object bounding boxes in each input frame, and FlowFormer [22] is

used to estimate the optical flow between consecutive frames. We standardize RGB, optical flow, and ego-motion inputs using means and standard deviations of input modalities on the training dataset. Note that the estimated homography matrices are normalized so that the element in the third row and the third column is one before standardization.

Network architecture. We use the dimension size of a token $C = 512$, $k = 2$ for the top-k confidence score with the threshold of 0.5, and set the number of blocks in the encoder and decoder to 2. Each block has 8 attention heads in the encoder and decoder. Our MLPs for hand and ego-motion prediction consist of a linear layer, an activation function of ReLU [1], a Dropout [46] layer, and a final linear layer that outputs the hand positions and ego-motion at future frames.

Optimization. We train the model for 30 epochs using the AdamW optimizer [34], with a peak learning rate of $2e-4$, linearly increased for the first 5 epochs of the training and decreased to 0.0 until the end of training with cosine decay [33]. We use weight decay of $1e-3$ and a batch size of 64. Regarding the parameters for the loss function, we empirically adapt the control point $\beta = 5.0$ in Eq. (6), and the loss weight of α, used in Eq. (9), is set to 1.

4.3   Evaluation Metrics

The distance between the predicted and ground truth positions in 2D image space, measured in pixels, is used to evaluate future hand position prediction performance. Specifically, we adopt traditional metrics of trajectory prediction [2, 6, 18]: average displacement error and final displacement error. Note that the metric is calculated using an image height scale of 256 px.

Average Displacement Error (ADE). ADE is calculated as the l_2 distance between the predicted future hand positions and the ground truth positions in pixel averaged over the entire future time steps and both left and right hands.

Final Displacement Error (FDE). FDE measures the l_2 distance between the predicted future hand positions and ground truth positions at the last time step and is averaged over two hands.

4.4   Comparison Methods

We compare with the following methods:

- **CVM** [43]. The Constant Velocity Model (CVM) is a simple but effective trajectory prediction method based on the assumption that the most recent relative motion is the most relevant predictor for the future trajectory. We compute the velocity (v_x, v_y) between $t = T - 1$ and $t = T$ for each hand (right, left), and future hand positions for $t = \{T+1, ..., T+F\}$ are forecasted using (v_x, v_y).
- **KF** [23]. The Kalman Filter is an algorithm for estimating a dynamic system's state based on noisy measurements. It tracks the center of the bounding boxes of the hands with its scale and aspect ratio. Our implementation is

Table 1: Intra-dataset evaluation. We assess the performance of future hand forecasting on two large-scale egocentric datasets, Ego4D and EPIC-Kitchens 55. In terms of input modalities, the symbols T_h, T_o, G_r, G_f, E represents *trajectory information* of hands and objects, *global information* of RGB and optical flow, and *ego-motion information*, respectively. Note that no backbone is used in CVM, KF, and Seq2Seq as these methods predict based on past trajectories and do not input RGB or optical flow frames. The best values are shown in **bold**, and the second best values are shown with underline.

Method	Input Modality	Backbone	Ego4D		EPIC-Kitchens 55	
			ADE ↓	FDE ↓	ADE ↓	FDE ↓
CVM [43]	T_h	-	108.11	143.23	141.70	155.40
KF [23]	T_h	-	71.23	72.87	70.58	75.60
Seq2Seq [47]	T_h	-	55.91	60.72	62.24	67.85
OCT [31]	T_h, T_o, G_r	BN-Inception	49.40	54.73	53.85	59.06
I3D + Regression [16]	G_r	3D ResNet-50	49.27	53.04	49.64	54.83
Ours	T_h, T_o, G_r, G_f, E	2D ResNet-18	**48.99**	**52.83**	**48.78**	**54.03**

based on the code provided by SORT [3][3], which adopts a Kalman Filter to track the center of bounding boxes.

- **Seq2Seq** [47]. Seq2Seq employs Long Short-Term Memory (LSTM) [20] to encode temporal information in the observation sequence and decode the target location of the hands. In our implementation, we adopt the embedding size of 512, the hidden dimension of 256, and the teacher forcing ratio of 0.5 during training.
- **OCT** [31]. OCT simultaneously predicts contact points and the hand trajectory. It takes RGB features extracted by BNInception [48], bounding boxes of hands and objects, and their cropped visual features as input. We modified the model not to predict the contact point for a fair comparison. Our implementation of this model is based on the official implementation[4].
- **I3D + Regression** [16]. This method is proposed as a benchmark for hand forecasting in the Ego4D dataset. The model is trained with the official hand forecasting code[5].

The first two traditional approaches predict based only on past trajectories without training. On the other hand, the last three methods above are recent advanced learning-based approaches in the hand forecasting task.

4.5   Hand Forecasting Accuracy Comparison

Intra-dataset evaluation. We compare the performance of hand forecasting with the prior methods on two large-scale egocentric datasets. Tab. 1 shows
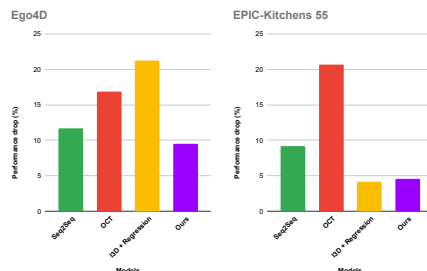
---

[3] `https://github.com/abewley/sort`

[4] `https://github.com/stevenlsw/hoi-forecast`

[5] `https://github.com/EGO4D/forecasting`

Table 2: Cross-dataset evaluation. $A \to B$ in the first row indicates that the models are trained on the training set of dataset A and tested on the validation set of dataset B. We conduct two cross-dataset evaluations: (1) trained on EPIC-Kitchens 55 and evaluated on Ego4D and (2) trained on Ego4D and evaluated on EPIC-Kitchens 55.

Method	EPIC → Ego4D		Ego4D → EPIC	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓
CVM [43]	108.11	143.23	141.70	155.40
KF [23]	71.23	72.87	70.58	75.60
Seq2Seq [47]	62.43	67.85	67.97	72.26
OCT [31]	<u>57.74</u>	<u>59.10</u>	64.97	65.84
I3D + Regression [16]	59.72	61.72	<u>51.70</u>	<u>58.37</u>
Ours	**53.67**	**56.36**	**51.03**	**56.78**

that the proposed method consistently outperforms the state-of-the-art methods. Our proposed method surpasses OCT by 9.4% (from 53.85 to 48.78) and I3D + Regression by 1.7% (from 49.64 to 48.78) on the EPIC-Kitchens 55 dataset. On the Ego4D dataset, our method exhibits similar performance on EPIC-Kitchens 55 and outperforms the prior works. Furthermore, the poor performance of the constant velocity model [43], which outperforms the learning-based approaches [18, 42] for pedestrian trajectory prediction from exocentric videos, confirms that the 2D hand forecasting task from egocentric videos presents unique challenges due to ego-motion.

Cross-dataset evaluation. We compare the generalization performance for future hand forecasting with the state-of-the-art methods in the cross-dataset scenario, where the domain of the test data is different from the training dataset. Tab. 2 summarizes the generalization performance of the comparison methods and the proposed method. Our proposed method surpasses OCT by 7.0% on the Ego4D dataset, where the models are trained on the EPIC-Kitchens 55 dataset.

Moreover, the learning-based approaches (OCT, I3D+Regression, and Ours) demonstrate lower accuracy in the cross-dataset scenario as compared to intra-dataset evaluations (see Fig. 3). This performance decrease stems from dataset bias, as the two

Fig. 3: The accuracy drop comparison. The figure summarizes the accuracy drop percentage in the cross-dataset scenario from the accuracy in the intra-dataset scenario for each method. A lower value indicates that the performance does not drop by changing the scenario from intra-dataset to cross-dataset. We summarize the performance drop of the learning-based model as there is no performance degradation in non-learnable methods, such as CVM and KF.

datasets originate from different distributions. The performance of I3D + Regression drops significantly (21.1%) when the model is trained on EPIC-Kitchens

Table 3: Action category-level evaluation. We compare the hand forecasting performance in the cross-dataset scenarios at the action category level with the conventional learning-based approaches. The results of five action categories, such as cooking, mechanic, arts/crafts, building, and gardening/farming, are summarized in the table.

Method	Cooking		Mechanic		Arts and crafts		Building		Gardening/farming	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
Seq2Seq [47]	58.45	60.73	59.78	62.83	64.60	66.85	68.28	70.11	64.42	66.52
OCT [31]	52.45	54.57	53.63	55.24	62.52	64.19	63.06	63.83	57.49	58.25
I3D + Regression [16]	48.26	52.26	58.03	59.73	63.03	64.89	67.55	68.83	61.80	62.98
Ours	**47.32**	**51.33**	**47.53**	**51.02**	**58.89**	**61.28**	**59.83**	**62.30**	**53.16**	**55.68**

55 and tested on Ego4D. On the other hand, although the accuracy is dropped in our method on Ego4D (9.6% dropping), the decrease is relatively small compared to other learning-based methods, thereby verifying the generalizability of the proposed method.

Action category-level evaluation. We conduct action category-level evaluations in the cross-dataset scenario, where the models are trained on EPIC-Kitchens 55 and tested on each action category on Ego4D to assess the generalizability among unseen actions. We focus on five major action categories on the Ego4D validation set: cooking, mechanic, arts/crafts, building, and gardening/farming. Tab. 3 demonstrates that our proposed method outperforms the prior learning-based methods across all categories. This indicates that our proposed method is highly generalizable to unseen action categories. In contrast, although the I3D + Regression method performs well in the cooking category, which is included in the training dataset, a significant performance gap can be seen in other categories compared to the cooking category. This occurs because I3D + Regression tends to overfit to the context and background of the training data, particularly in the cooking category.

4.6   Ablation Analysis

Input modality. The ablation study focuses on the input modalities to verify the contribution of each input component to the overall performance in intra/cross-dataset settings. We experiment by removing each input modality: bounding boxes of objects, RGB frame, optical flow, and ego-motion information. As shown in Tab. 4, the absence of visual or flow information degrades the performance by 2.4% (from 48.89 to 50.08) and 4.3% (from 48.89 to 51.00) on intra-dataset evaluation on average, respectively.

Moreover, although the absence of object or ego-motion information outperforms the proposed method on intra-dataset evaluation, these methods degrade the prediction performance on cross-dataset scenarios. This performance deterioration on cross-dataset scenarios indicates that leveraging all input modalities (the proposed method), including ego-motion information, is beneficial for unseen scenes.

Table 4: Input modality ablation study. Ablation study on the input modalities on Ego4D and EPIC-Kitchens 55. We summarize the results of two scenarios, intra or cross-dataset. The last column is the result of the proposed method, which uses all the modal information.

Table 5: Loss component ablation study. Ablation study on ego-motion estimation loss on the two datasets in intra and cross-dataset scenarios to verify the effectiveness of propagating ego-motion estimation loss.

Object	RGB	Flow	Ego	Intra		Cross	
				ADE ↓	FDE ↓	ADE ↓	FDE ↓
	✓	✓	✓	<u>48.76</u>	53.79	<u>52.78</u>	<u>57.02</u>
✓		✓	✓	50.08	54.83	53.30	57.54
✓	✓		✓	51.00	54.78	54,74	57.93
✓	✓	✓		**48.35**	**53.24**	52.89	<u>57.02</u>
✓	✓	✓	✓	48.89	<u>53.43</u>	**52.35**	**56.57**

| Method	Intra		Cross	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓
w/o $\mathcal{L}_{ego}$	49.66	54.26	52.84	57.08
w/ $\mathcal{L}_{ego}$ (Ours)	**48.89**	**53.43**	**52.35**	**56.57**

Loss. We also perform an ablation study on the loss function. We evaluate the advantage of the ego-motion estimation loss term $\mathcal{L}_{ego}$ in Eq. (9). Tab. 5 shows that training the proposed method without the ego-motion estimation loss $\mathcal{L}_{ego}$ deteriorates hand forecasting performance by 1.6% and 0.9% in terms of ADE in the intra/cross-dataset scenario, respectively. This degradation verifies the effectiveness of the proposed method, which forecasts the camera wearer's future ego-motion as an auxiliary task.

4.7   Qualitative Results

The qualitative results on the Ego4D and EPIC-Kitchens 55 datasets are visualized in Fig. 4. We present two sequences from EPIC-Kitchens 55 in the top two rows of the figure and two sequences from Ego4D in the bottom two rows. In the second sequence from the top of EPIC-Kitchens, where the camera wearer turns left, the proposed method predicts the hand positions more accurately than the other methods. This capability of prediction, even in the presence of ego-motion, verifies the effectiveness of our ego-motion-aware model.

5   Conclusion

Conclusion. We present EMAG, the first model to explore the potential benefit of incorporating ego-motion into the hand forecasting task. We propose leveraging the homography matrix to represent the camera wearer's ego-motion and to verify its effectiveness. Furthermore, our proposed method utilizes multiple modalities to mitigate the susceptibility to overfitting to backgrounds or scene textures. Experiments on two large-scale egocentric datasets, Ego4D and EPIC-Kitchens 55, demonstrate that our simple but effective approach outperforms
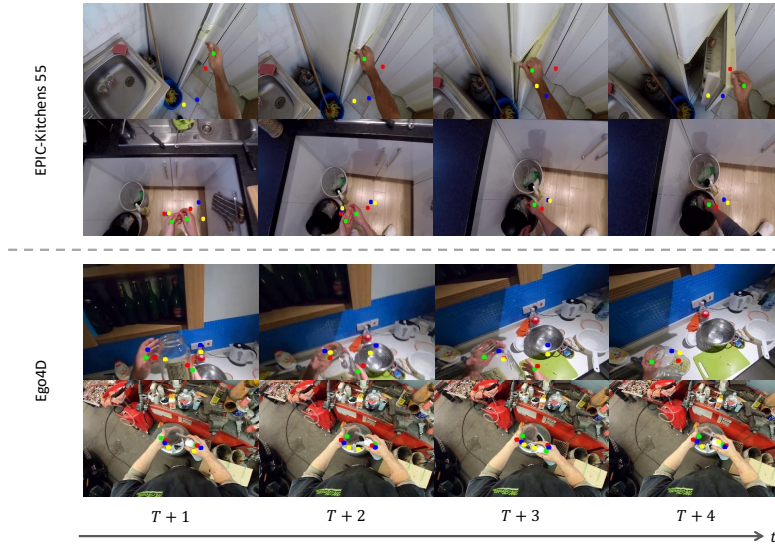
Fig. 4: Qualitative results. We present two sequences of predictions each from Ego4D and EPIC-Kitchens 55. Dots colored in green, red, blue, and yellow represent the hand positions of the ground truth, the proposed method, I3D + Regression, and OCT, respectively.

the state-of-the-art hand forecasting methods in terms of accuracy and generalizability against unseen scenes and actions.

Limitations and future work. Our proposed method leverages the trajectory information of hands and objects detected based on the off-the-shelf hand object detector [44] from egocentric video. Thus, the bias and errors from the off-the-shelf detector may still affect the input trajectory information. In addition, the proposed method requires multiple pre-processing modules, such as hand object detection, optical flow estimation, and homography matrix estimation. However, efficient and real-time inference capabilities on edge devices are essential for forecasting in real-world applications. We will leave this for our future efforts.

Acknowledgements

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018)

2. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: Proceedings of the IEEE International Conference on Image Processing (ICIP) (2016)

4. Birkfellner, W., Figl, M., Huber, K., Watzinger, F., Wanschitz, F., Hummel, J., Hanel, R., Greimel, W., Homolka, P., Ewers, R., Bergmann, H.: A head-mounted operating binocular for augmented reality visualization in medicine - design and initial evaluation. IEEE Transactions on Medical Imaging (TMI) **21**(8), 991–997 (2002)

5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

6. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

7. Choi, C., Kim, S.M., Kim, Y.M.: Balanced spherical grid for egocentric view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

9. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision (IJCV) **130**(1), 33–55 (2022)

10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)

11. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)

12. Fotouhi, J., Mehrfard, A., Song, T., Johnson, A., Osgood, G., Unberath, M., Armand, M., Navab, N.: Development and pre-clinical analysis of spatiotemporal-aware augmented reality in orthopedic interventions. IEEE Transactions on Medical Imaging (TMI) **40**(2), 765–778 (2021)

13. Fu, C.Y., Shvets, M., Berg, A.C.: Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. arXiv preprint arXiv:1901.03353 (2019)

14. Furnari, A., Farinella, G.M.: Rolling-unrolling lstms for action anticipation from first-person video. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **43**(11), 4021–4036 (2020)

15. Gong, X., Mohan, S., Dhingra, N., Bazin, J.C., Li, Y., Wang, Z., Ranjan, R.: Mmg-ego4d: Multimodal generalization in egocentric action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
17. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
18. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
21. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
22. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: FlowFormer: A transformer architecture for optical flow. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
23. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82**(1), 35–45 (1960)
24. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
25. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning cross-modal contrastive features for video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
26. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
27. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

28. Li, Y., Cao, Z., Liang, A., Liang, B., Chen, L., Zhao, H., Feng, C.: Egocentric prediction of action target in 3d. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

29. Liu, M., Ma, L., Somasundaram, K., Li, Y., Grauman, K., Rehg, J.M., Li, C.: Egocentric activity recognition and localization on a 3d map. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)

30. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In: Proceddings of the European Conference on Computer Vision (ECCV) (2020)

31. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

32. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

33. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: Proceedings of the International Conference on Learning Representations (ICLR) (2017)

34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)

35. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

36. Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., Caputo, B.: E2(go)motion: Motion augmented event stream for egocentric action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

37. Price, W., Vondrick, C., Damen, D.: Unweavenet: Unweaving activity stories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

38. Qi, Z., Wang, S., Su, C., Su, L., Huang, Q., Tian, Q.: Self-regulated learning for egocentric video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **45**(6), 6715–6730 (2023)

39. Quintero, C.P., Li, S., Pan, M.K., Chan, W.P., Machiel Van der Loos, H., Croft, E.: Robot programming through augmented trajectories in augmented reality. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2018)

40. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Spotem: Efficient video search for episodic memory. In: International Conference on Machine Learning (ICML) (2023)

41. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

42. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

43. Schöller, C., Aravantinos, V., Lay, F., Knoll, A.: What the constant velocity model can teach us about pedestrian motion prediction. IEEE Robotics and Automation Letters (RA-L) **5**(2), 1696–1703 (2020)
44. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
45. Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., Zisserman, A.: A short note on the kinetics-700-2020 human action dataset. arXiv preprint arXiv:2010.10864 (2020)
46. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research (JMLR) **15**(56), 1929–1958 (2014)
47. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2014)
48. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
49. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2017)
51. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
52. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
53. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
54. Weinzaepfel, P., Rogez, G.: Mimetics: Towards understanding human actions out of context. International Journal of Computer Vision (IJCV) **129**(5), 1675–1690 (2021)
55. Whitney, J.P., Chen, T., Mars, J., Hodgins, J.K.: A hybrid hydrostatic transmission and human-safe haptic telepresence robot. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2016)
56. Wilmott, J.P., Erkelens, I.M., Murdison, T.S., Rio, K.W.: Perceptibility of jitter in augmented reality head-mounted displays. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (2022)
57. Xu, W., Liang, H.N., He, A., Wang, Z.: Pointing and selection methods for text entry in augmented reality head mounted displays. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (2019)
58. Xue, Z., Song, Y., Grauman, K., Torresani, L.: Egocentric video task translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
59. Yu, Z., Huang, Y., Furuta, R., Yagi, T., Goutsu, Y., Sato, Y.: Fine-grained affordance annotation for egocentric hand-object interaction videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023)