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

Masashi Hatano[1]
hatano1210@keio.jp
Ryo Hachiuma[1,2]
ryo-hachiuma@keio.jp
Hideo Saito[1]
hs@keio.jp

[1] Keio University
Yokohama, Japan
[2] NVIDIA
Taipei, Taiwan

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 7.0% on cross-dataset evaluations. Project page: https://masashi-hatano.github.io/EMAG/

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 [6, 7] and Ego4D [14] 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 [13, 31, 46], human body pose estimation [23, 44, 45], audio-visual understanding [18, 36], action anticipation [12, 33], and natural language queries [35].

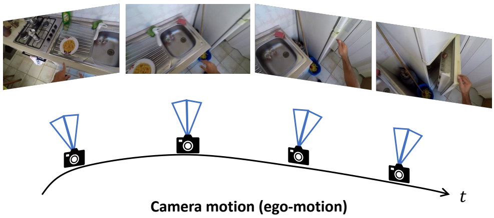

Figure 1: The presence of ego-motion in first-person videos significantly affects the dynamic movement of the camera position. Since the camera is part of the wearer's body, a variety of views can be captured even in a short period of time.

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 [49, 50] and human-robot interactions [34, 48] 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 [14, 26, 28] 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 [14] uses a simple I3D network [3] and regresses the future 2D hand coordinates. Meanwhile, the Object Centric Transformer (OCT) [28] is a method that jointly predicts hand motions and object contact points from RGB video features extracted with BNInception [41] 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 [22, 47]. 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 predicting 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 [31]. Consequently, the model's generalizability is enhanced, ensuring robust performance even when distribution gaps exist between the training and test data.

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 on cross-dataset scenarios in which the training and test datasets differ.

2 Related Work

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 [15, 21, 39] 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 [5, 24, 27, 30, 32, 51]. 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 [49, 50] or medical image analysis [2, 10].

Multiple large-scale egocentric video datasets [6, 7, 14, 25, 29] 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 [13, 31, 46], human-object interaction [26, 28, 52], action anticipation [12, 33], human body pose estimation [23, 44, 45], and audio-visual understanding [18, 36]. In this work, we explore one of the challenging tasks in egocentric video analysis, 2D hand forecasting.

Hand Forecasting from Egocentric Videos. To predict future hand positions, traditional tracking or sequential methods, such as Kalman Filter (KF) [20], Constant Velocity Model (CVM) [37], and Seq2Seq [40], 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 [14] dataset, utilized I3D [3], 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 [26] is the first work to incorporate the future trajectory of hands for action anticipation in egocentric videos. Building upon this, OCT [28] is an approach that integrates hand-object interactions into the prediction process.

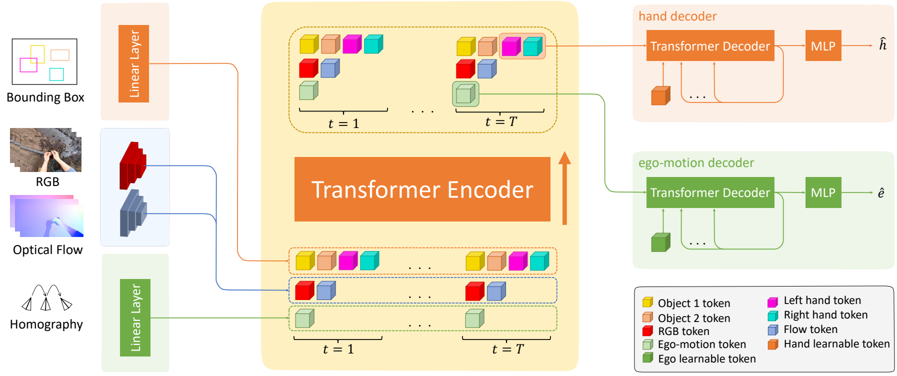

Figure 2: The architecture of the proposed method. Given input egocentric video frames, we pre-process them and obtain multiple modalities, including RGB and optical flow, detected bounding boxes of objects/hands, and homography matrices of adjacent frames. We train a single Transformer encoder and two Transformer decoders with MLP heads for hand and ego-motion prediction.

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 [43]. 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.

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 [14]. 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 [38], 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 [42] or FlowFormer [19].

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 [9], 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 [43] 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 ($\mathcal{D}_{\text{hand}}$) and the ego-motion decoder ($\mathcal{D}_{\text{ego}}$), 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, $\mathcal{M}_{\text{hand}}$ and $\mathcal{M}_{\text{ego}}$, 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 $\boldsymbol{p}_{\text{hand}}$ for the query of the first forecasting time step ($\boldsymbol{q}_{\text{hand}}^T = \boldsymbol{p}_{\text{hand}}$):

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

where $\boldsymbol{q}_{\text{hand}}^{T+f}$ 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 $\boldsymbol{p}_{\text{ego}}$ ($= \boldsymbol{q}_{\text{ego}}^T$):

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

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. $\mathcal{M}_{\text{hand}}$ predicts the coordinates of the left and right hands $\hat{\boldsymbol{h}}^{T+f}$ at the future time step $T+f$. Similarly, $\mathcal{M}_{\text{ego}}$ predicts the nine elements of the homography matrix $\hat{\boldsymbol{e}}^{T+f}$:

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

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

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

3.5 Training Objective

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

Hand forecasting loss. We adopt the self-adjusting smooth L1 loss, which was introduced in RetinaMask [], 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 [6]. 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 RUL-STM [12].

Ego4D [14]. 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 [14].

Followed by the previous work [28], we employ the egocentric hand-object detector [38] 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 [17] on ImageNet [8] as the backbone to extract RGB and optical flow features. We adopt the hand and object detector from the egocentric video [38] to detect left/right hand and object bounding boxes in each input frame, and FlowFormer [19] 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.

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 [1, 4, 16]: average displacement error (ADE) and final displacement error (FDE). Note that the metric is calculated using an image height scale of 256 px. ADE is calculated as the l_2 distance between

Table 1: **Cross-dataset evaluation**. $A \rightarrow B$ 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. 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.

Method	Input Modality	Backbone	EPIC → Ego4D		Ego4D → EPIC	
			ADE ↓	FDE ↓	ADE ↓	FDE ↓
CVM [37]	T_h	-	108.11	143.23	141.70	155.40
KF [20]	T_h	-	71.23	72.87	70.58	75.60
Seq2Seq [40]	T_h	-	62.43	67.85	67.97	72.26
OCT [28]	T_h, T_o, G_r	BN-Inception	<u>57.74</u>	<u>59.10</u>	64.97	65.84
I3D + Regression [14]	G_r	3D ResNet-50	59.72	61.72	<u>51.70</u>	<u>58.37</u>
Ours	T_h, T_o, G_r, G_f, E	2D ResNet-18	**53.67**	**56.36**	**51.03**	**56.78**

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. 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.

4.3 Hand Forecasting Accuracy Comparison

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. 1 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.

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 (See Supp. A for further details). Tab. 2 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.

4.4 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 cross-dataset settings. We experiment by removing each input modality: bounding boxes of objects, RGB frame, optical flow, and

Table 2: **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 and farming	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
Seq2Seq [40]	58.45	60.73	59.78	62.83	64.60	66.85	68.28	70.11	64.42	66.52
OCT [28]	52.45	54.57	<u>53.63</u>	<u>55.24</u>	<u>62.52</u>	<u>64.19</u>	<u>63.06</u>	<u>63.83</u>	<u>57.49</u>	<u>58.25</u>
I3D + Regression [14]	<u>48.26</u>	<u>52.26</u>	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**

Table 3: **Input modality ablation study**. Ablation study on the input modalities on Ego4D and EPIC-Kitchens 55. The last column is the result of the proposed method, which uses all the modal information.

Object	RGB	Flow	Ego	ADE ↓	FDE ↓
	✓	✓	✓	<u>52.78</u>	<u>57.02</u>
✓		✓	✓	53.30	57.54
✓	✓		✓	54.74	57.93
✓	✓	✓		52.89	<u>57.02</u>
✓	✓	✓	✓	**52.35**	**56.57**

Table 4: **Loss component ablation study**. Ablation study on ego-motion estimation loss to verify the effectiveness of propagating ego-motion estimation loss.

Method	ADE ↓	FDE ↓
w/o $\mathcal{L}_{\text{ego}}$	52.84	57.08
w/ $\mathcal{L}_{\text{ego}}$ (Ours)	**52.35**	**56.57**

ego-motion information. As shown in Tab. 3, the absence of visual or flow information degrades the performance by 0.95% (from 52.35 to 53.30) and 2.4% (from 52.35 to 54.75) in terms of ADE on cross-dataset evaluation on average, respectively. Moreover, the absence of object or ego-motion information degrades 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.

Loss. We also perform an ablation study on the loss function. We evaluate the advantage of the ego-motion estimation loss term $\mathcal{L}_{\text{ego}}$ in Eq. (9). Tab. 4 shows that training the proposed method without the ego-motion estimation loss $\mathcal{L}_{\text{ego}}$ deteriorates hand forecasting performance by 0.9% in terms of ADE in the cross-dataset scenario. This degradation verifies the effectiveness of the proposed method, which forecasts the camera wearer's future ego-motion as an auxiliary task.

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 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 [38] 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.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] W. Birkfellner, M. Figl, K. Huber, F. Watzinger, F. Wanschitz, J. Hummel, R. Hanel, W. Greimel, P. Homolka, R. Ewers, and H. Bergmann. 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.

[3] Joao Carreira and Andrew Zisserman. 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.

[4] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Changwoon Choi, Sang Min Kim, and Young Min Kim. Balanced spherical grid for egocentric view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130(1):33–55, 2022.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Martin A. Fischler and Robert C. Bolles. 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.

[10] Javad Fotouhi, Arian Mehrfard, Tianyu Song, Alex Johnson, Greg Osgood, Mathias Unberath, Mehran Armand, and Nassir Navab. Development and pre-clinical analysis of spatiotemporal-aware augmented reality in orthopedic interventions. *IEEE Transactions on Medical Imaging (TMI)*, 40(2):765–778, 2021.

[11] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C. Berg. Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. *arXiv preprint arXiv:1901.03353*, 2019.

[12] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4021–4036, 2020.

[13] Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 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.

[15] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 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.

[16] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[20] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[22] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[23] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[24] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. 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.

[25] Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. 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.

[27] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M. Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[28] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. 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.

[29] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. 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.

[32] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[33] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(6):6715–6730, 2023.

[34] Camilo Perez Quintero, Sarah Li, Matthew KXJ Pan, Wesley P. Chan, H.F. Machiel Van der Loos, and Elizabeth Croft. Robot programming through augmented trajectories in augmented reality. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[35] Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Spotem: Efficient video search for episodic memory. In *International Conference on Machine Learning (ICML)*, 2023.

[36] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M. Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[37] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):1696–1703, 2020.

[38] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[39] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

[40] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[44] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[45] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[46] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[47] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision (IJCV)*, 129(5):1675–1690, 2021.

[48] John P. Whitney, Tianyao Chen, John Mars, and Jessica K. Hodgins. A hybrid hydrostatic transmission and human-safe haptic telepresence robot. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[49] James P. Wilmott, Ian M. Erkelens, T. Scott Murdison, and Kevin W. Rio. Perceptibility of jitter in augmented reality head-mounted displays. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2022.

[50] Wenge Xu, Hai-Ning Liang, Anqi He, and Zifan Wang. 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.

[51] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. Egocentric video task translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[52] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. 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.