Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video

Miao Liu[1], Siyu Tang[3], Yin Li[2], and James M. Rehg[1]

[1] Georgia Institute of Technology, Atlanta, United States
[2] University of Wisconsin-Madison, Madison, United States
[3] ETH Zürich, Switzerland

Abstract. We address the challenging task of anticipating human-object interaction in first person videos. Most existing methods either ignore how the camera wearer interacts with objects, or simply considers body motion as a separate modality. In contrast, we observe that the intentional hand movement reveals critical information about the future activity. Motivated by this observation, we adopt intentional hand movement as a feature representation, and propose a novel deep network that jointly models and predicts the egocentric hand motion, interaction hotspots and future action. Specifically, we consider the future hand motion as the motor attention, and model this attention using probabilistic variables in our deep model. The predicted motor attention is further used to select the discriminative spatial-temporal visual features for predicting actions and interaction hotspots. We present extensive experiments demonstrating the benefit of the proposed joint model. Importantly, our model produces new state-of-the-art results for action anticipation on both EGTEA Gaze+ and the EPIC-Kitchens datasets. Our project page is available at https://aptx4869lm.github.io/ForecastingHOI/

Keywords: First Person Vision, Action Anticipation, Motor Attention

1 Introduction

The human ability of "looking into the near future" remains a key challenge for computer vision. Consider the example in Fig. 1, given a video shortly before the start of an action, we can easily predict what will happen next, e.g., the person will take the canister of salt. Even without seeing any future frames, we can vividly imagine how the person will perform the action, e.g., the trajectory of the hand when reaching for the canister or the location on the canister that will be grasped.

There is convincing evidence that our remarkable ability to forecast other individuals' actions depends critically upon our perception and interpretation of their body motion. The investigation of this anticipatory mechanism dates back to 19th century, when William James argued that future expectations are intrinsically related to purposive body movements [25]. Additional evidence for
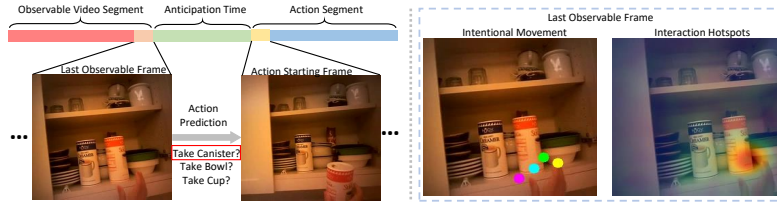
Fig. 1. *What is the most likely future interaction?* Our model takes advantage of the connection between motor attention and visual perception. In addition to future action label, our model also predicts the interaction hotspots on the last observable frame and hand trajectory (in the order of yellow , green, cyan, and magenta) between the last observable time step to action starting point. Visualizations of hand trajectory are projected to the last observable frame (best viewed in color).

a link between perceiving and performing actions was provided by the discovery of mirror neurons [8,20]. The observation of others' actions activates our motor cortex, the same brain regions that are in charge of the planning and control of intentional body motion. This activation can happen even before the onset of the action and is highly correlated with the anticipation accuracy [1]. A compelling explanation from [45] suggests that *motor attention*, i.e., the active prediction of meaningful future body movements, serves as a key representation for anticipation. A goal of this work is to develop a computational model for motor attention that can enable more accurate action prediction.

Despite these relevant findings in cognitive neuroscience, the role of intentional body motion in action anticipation is largely ignored by the existing literature [56,11,16,27,13,15,38,28]. In this work, we focus on the problem of forecasting human-object interactions in First Person Vision (FPV). Interactions consist of a single verb and one or more nouns, with "take bowl" as an example. FPV videos capture complex hand movements during a rich set of interactions, thus providing a powerful vehicle for studying the connection between motor attention and future representation. Several previous works have investigated the problems of FPV activity anticipation [13,15] and body movement prediction [2,19,12,57]. We believe we are the first to utilize a motor attention model for FPV action anticipation.

To this end, we propose a novel deep model that predicts "motor attention"— the future trajectory of the hands, as an anticipatory representation of actions. Based on motor attention, our model further localizes the future contact region of the interaction, i.e., interaction hotspots [39] and recognizes the type of future interactions. Importantly, we characterize motor attention and interaction hotspots as probabilistic variables modeled by stochastic units in a deep network. These units naturally deal with the uncertainty of future hand motion and contact region during interaction, and produce attention maps that highlight discriminative spatial-temporal features for action anticipation.

During inference, our model takes video clips shortly before the interaction as inputs, and jointly predicts motor attention, interaction hotspots, and action

labels. During training, our model assumes that these outputs are available as supervisory signals. To evaluate our model, we report results on two major FPV benchmarks: EGTEA Gaze+ and EPIC-Kitchens. Our approach outperforms prior state-of-the-art methods by a significant margin. In addition, we conduct extensive ablation studies to verify the design of our model and evaluate our model for motor attention prediction and interaction hotspots estimation. Our model demonstrates strong results for both tasks. We believe our model provides a solid step towards the challenge of FPV visual anticipation.

2   Related Works

There has recently been substantial interest in learning to forecast future events in videos. The most relevant works to ours are those investigations on FPV action anticipation. Our work is also related to previous studies on third person action anticipation, other visual prediction tasks, and visual affordance.

FPV Action Anticipation. Action anticipation aims at predicting an action before it happens. We refer the readers to a recent survey [30] for a distinction between action recognition and anticipation. FPV action recognition has been studied extensively [46,42,10,63,36,34,32,41], while fewer works have targeted egocentric action anticipation. Shen et al. [49] investigated how different egocentric modalities affect the action anticipation performance. Soran et al. [52] adopted Hidden Markov Model to compute the transition probability among sequences of actions. A similar idea was explored in [38]. Furnari et al. [13] considered the task of predicting the next-active objects. Their recent work [15] proposed to factorize the anticipation model into a "Rolling" LSTM that summarizes the past activity and an "Unrolling" LSTM that makes hypotheses of the future activity. Ke et al. [28] proposed a time-conditioned skip connection operation to extract relevant information for action anticipation. In contrast to our proposed method, these prior works did not exploit the connection between human motor attention and visual perception, and did not explicitly model the contact region during human-object interaction.

Third Person Action Anticipation. Several previous efforts seek to address the task of action anticipation in third person vision. Kris et al. [29] combined semantic scene labeling with a Markov decision process to forecast the behavior and trajectory of a subject. Vondrick et al. [56] proposed to predict the future video representation from large scale unlabeled video data. Gao et al. [16] proposed a Reinforced Encoder-Decoder network to create a summary representation of past frames and produce a hypothesis of future action. Kataoka et al. [27] introduced a subtle motion descriptor to identify the difference between an on-going action and a transitional action, and thereby facilitate future anticipation. Our work shares the same goal of future forecasting, but we focus on leveraging abundant visual cues from egocentric videos for action anticipation.

Other Prediction Tasks. Anticipation has been studied under other vision tasks. In particular, human body motion prediction has been extensively studied [40,55,58,19,57,12], including recent work in the setting of FPV. Rhinehart et

al. [44] proposed an online learning algorithm to forecast the first-person trajectory. Park et al. [51] proposed a deep network to infer possible human trajectories from egocentric stereo images. Wei et al. [61] utilized a probabilistic model to infer 3D human attention and intention. Tagi et al. [62] addressed a novel task of predicting the future locations of an observed subject in egocentric videos. Ryoo et al. [47] proposed a novel method to summarize pre-activity observations for robot-centric activity prediction. However, none of these previous work considered modeling body movement for action anticipation.

Visual Affordance. The problem of predicting visual affordances has attracted growing interest in computer vision. Affordance can be helpful for scene understanding [18,7,59], human-object interaction recognition [53], and action analysis [43,31]. Several recent works have focused on estimating visual affordances that are grounded on human object interaction. Chen et al. [5] proposed to estimate likely object interaction regions by learning the connection between subject and object. Fang et al. [9] proposed to estimate interaction regions by learning from demonstration videos. However, none of these previous works considered future prediction. More recently, Tushar et al. [39] introduced an unsupervised learning method that uses the backward attention map to approximate the interaction hotspots grounded on a future action. However, their method did not model the presence of objects and thus can not be used to anticipate human-object interactions. However, we compare to their results for interaction hotspot estimation in our experiments.

3   Method

We consider the setting of action anticipation from [6]. Denote an input video segment as $x : [\tau_a - \Delta\tau_o, \tau_a]$. x starts at $\tau_a - \Delta\tau_o$ and ends at τ_a with duration $\Delta\tau_o > 0$ as the "observation time". Our goal is to predict the label y of an immediate future interaction starting at $\tau_s = \tau_a + \Delta\tau_a$, where $\Delta\tau_a > 0$ is a fixed interval known as the "anticipation time." Moreover, we seek to estimate future hand trajectories $\mathcal{M}$ within $[\tau_a, \tau_s]$ (projected back to the last observable frame at τ_a), and to localize interaction hotspots $\mathcal{A}$ at τ_a (the last observable frame). Fig. 1 illustrates our setting.

To summarize, our model seeks to anticipate the future action y by jointly predicting the future hand trajectory $\mathcal{M}$ and interaction hotspots $\mathcal{A}$ at the last observable frame. Predicting the future is fundamentally ambiguous, since the observation of future interaction only represents one of the many possibilities characterized by an underlying distribution. Our key idea is thus to model motor attention and interaction hotspots as probabilistic variables in order to account for their uncertainty. We present an overview of our model in Fig. 2.

Specifically, we make use of a 3D backbone network $\phi(x)$ for video representation learning. Following the approach in [50,21], we utilize 5 convolutional blocks, and denote the features from the ith convolution block as $\phi_i(x)$. Based on $\phi(x)$, our motor attention module (b) predicts future hand trajectories as motor attention $\mathcal{M}$ and uses stochastic units to sample from $\mathcal{M}$. The sampled motor
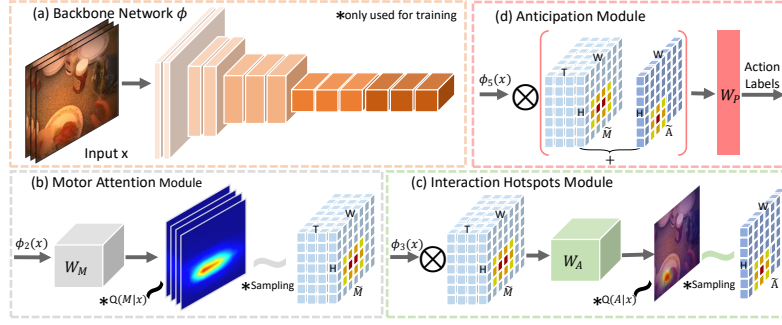
Fig. 2. Overview of our model. A 3D convolutional network $\phi(x)$ is used as our backbone network, with features from its ith convolution block as $\phi_i(x)$ (a). A motor attention module (b) makes use of stochastic units to generate sampled future hand trajectories $\tilde{\mathcal{M}}$ used to guide interaction hotspots estimation in module (c). Module (c) further generates sampled interaction hotspots $\tilde{\mathcal{A}}$ with a similar stochastic units as in module (b). Both $\tilde{\mathcal{M}}$ and $\tilde{\mathcal{A}}$ are used to guide action anticipation in anticipation module (d). During testing, our model takes only video clips as inputs, and predicts motor attention, interaction hotspots, and action labels. Note that $\otimes$ represents element-wise multiplication for weighted pooling.

attention $\tilde{\mathcal{M}}$ is an indicator of important spatial-temporal features for interaction hotspot estimation. Our interaction hotspot module (c) further produces an interaction hotspot distribution $\mathcal{A}$ and its sample $\tilde{\mathcal{A}}$. Finally, our anticipation module (d) makes use of both $\tilde{\mathcal{M}}$ and $\tilde{\mathcal{A}}$ to aggregate network features, and predicts the future interaction y.

3.1   Joint Modeling of Human-Object Interaction

Formally, we consider motor attention $\mathcal{M}$ and interaction hotspots $\mathcal{A}$ as probabilistic variables, and model the conditional probability of the future action label y given the input video x as a latent variable model, where

$$p(y|x) = \int_{\mathcal{M}} \int_{\mathcal{A}} p(y|\mathcal{A}, \mathcal{M}, x) p(\mathcal{A}|\mathcal{M}, x) p(\mathcal{M}|x) \ d\mathcal{A} \ d\mathcal{M}, \tag{1}$$

$p(\mathcal{M}|x)$ first estimates motor attention from video input x. $\mathcal{M}$ is further used to estimate interaction hotspots A ($p(\mathcal{A}|\mathcal{M}, x)$). Given x, $\mathcal{M}$ and $\mathcal{A}$, the action label y is determined by $p(y|\mathcal{A}, \mathcal{M}, x)$. Our model thus consists of three main components.

Motor Attention Module tackles $p(\mathcal{M}|x)$. Given the network features $\phi_2(x)$, our model uses a function F_M to predict motor attention $\mathcal{M}$. $\mathcal{M}$ is represented as a 3D tensor of size $T_m \times H_m \times W_m$. Moreover, $\mathcal{M}$ is normalized within each temporal slice, i.e., $\sum_{w,h} \mathcal{M}(t, w, h) = 1$.

Interaction Hotspots Module targets at $p(\mathcal{A}|\mathcal{M}, x)$. Our model uses a function F_A to estimate the interaction hotspots $\mathcal{A}$ based on the network feature

$\phi_3(x)$ and sampled motor attention $\tilde{\mathcal{M}}$. $\mathcal{A}$ is represented as a 2D attention map of size $H_a \times W_a$. A further normalization constrained that $\sum_{w,h} \mathcal{A}(w, h) = 1$.

Anticipation Module makes use of the predicted motor attention and interaction hotspots for action anticipation. Specifically, sampled motor attention $\tilde{\mathcal{M}}$ and sampled interaction hotspots $\tilde{\mathcal{A}}$ are used to aggregate feature $\phi_5(x)$ via weighted pooling. An action anticipation function F_P further maps the aggregated features to future action label y.

3.2   Motor Attention Module

Motor Attention Generation. The motor attention prediction function F_M is composed of a linear function with parameter W_M on top of network features $\phi_2(x)$. The linear function is realized by a 3D convolution and a softmax function is used to normalized the attention map. This is given by $\psi = softmax(W_M^T \phi_2(x))$, where the output ψ is a 3D tensor of size $T_m \times H_m \times W_m$. We further model $p(\mathcal{M}|x)$ by normalizing ψ within each temporal slice:

$$\mathcal{M}_{m,n,t} = \frac{\psi_{m,n,t}}{\sum_{m,n} \psi_{m,n,t}}, \tag{2}$$

where $\psi_{m,n,t}$ is the value at location (m, n) and time step t in the 3D tensor of ψ. And $\mathcal{M}$ can be considered as the expectation of $p(\mathcal{M}|x)$.

Stochastic Modeling. Modeling motor attention in the context of forecasting human-object interaction requires a mechanism for addressing the stochastic nature of motor attention in developing the joint model. Here, we propose to use stochastic units to model the uncertainty. The key idea is to sample from the motor attention distribution. We follow the Gumbel-Softmax and reparameterization trick introduced in [26,37] to design a differentiable sampling mechanism:

$$\tilde{\mathcal{M}}_{m,n,t} \sim \frac{\exp((\log \psi_{m,n,t} + G_{m,n,t})/\theta)}{\sum_{m,n} \exp((\log \psi_{m,n,t} + G_{m,n,t})/\theta)}, \tag{3}$$

where G is a Gumbel Distribution used to sample from discrete distribution. This Gumbel-Softmax trick produces a "soft" sampling step that allows the direct back-propagation of gradients to ψ. θ is the temperature parameter that controls the "sharpness" of the distribution. We set $\theta = 2$ for all of our experiments.

3.3   Interaction Hotspots Module

The predicted motor attention $\mathcal{M}$ is further used to guide interaction hotspots estimation $p(\mathcal{A}|x)$ by considering the conditional probability

$$p(\mathcal{A}|x) = \int_{\mathcal{M}} p(\mathcal{A}|\mathcal{M}, x) p(\mathcal{M}|x) d\mathcal{M}. \tag{4}$$

In practice, $p(\mathcal{A}|x)$ is estimated using sampled motor attention $\tilde{\mathcal{M}}$ based on $p(\mathcal{A}|\tilde{\mathcal{M}}, x)$ and $p(\tilde{\mathcal{M}}|x)$. For each sample $\tilde{\mathcal{M}}$, $p(\mathcal{A}|\tilde{\mathcal{M}}, x)$ is defined by the interaction hotspots estimation function F_A. F_A takes the input of a motor attention

map $\tilde{\mathcal{M}}$ and $\phi_3(x)$, and has the form of a linear 2D convolution parameterzied by W_A followed by a softmax function.

$$p(\mathcal{A}|\tilde{\mathcal{M}}, x) = softmax\left(W_A^T(\tilde{\mathcal{M}} \otimes \phi_3(x))\right), \tag{5}$$

where $\otimes$ is the Hadamard product (element-wise multiplication). The result $p(\mathcal{A}|\mathcal{M}, x)$ is a 2D map of size $H_a \times W_a$. Intuitively, $\tilde{\mathcal{M}}$ presents a spatial-temporal saliency map to highlight feature representation $\phi_3(x)$. F_A thus normalizes (using softmax) the output of a linear model on the selected features $\tilde{\mathcal{M}} \otimes \phi_3(x)$, and is a convex function. Finally, a similar sampling mechanism as in Eq. 3 can be used to sample $\tilde{\mathcal{A}}$ from $p(\mathcal{A}|x)$.

3.4 Anticipation Module

We now present the last piece of our model—the action anticipation module. The action anticipation function $p(y|\mathcal{A}, \mathcal{M}, x) = F_P(\mathcal{A}, \mathcal{M}, x)$ is defined as a function of the sampled motor attention map (3D) $\tilde{\mathcal{M}}$, sampled interaction heatmap (2D) $\tilde{\mathcal{A}}$ and the network feature $\phi_5(x)$. This is given by

$$p(y|\tilde{\mathcal{A}}, \tilde{\mathcal{M}}, x) = softmax\left(W_P^T \Sigma\left(\tilde{\mathcal{M}} \otimes \phi_5(x)\right) + W_P^T \Sigma\left(\tilde{\mathcal{A}} \odot \phi_5(x)\right)\right), \tag{6}$$

where $\otimes$ is again the Hadamard product. Σ is the global average pooling operation that pools a vector representation from a 2D or 3D feature map. $\odot$ is to use a 2D map ($\tilde{\mathcal{A}}$) to conduct Hadamard product to the last temporal slice of a 3D tensor $\phi_5(x)$. This is because the interaction hotspots $\tilde{\mathcal{A}}$ is only defined on the last observable frame. W_P is a linear function that maps the features into prediction logits. F_P is a combination of linear operations followed by a softmax function, and thus remains a convex function.

3.5 Training and Inference

Training our proposed joint model is challenging, as $p(\mathcal{M}|x)$ and $p(\mathcal{A}|\mathcal{M}, x)$ are intractable. Fortunately, variational inference comes to the rescue.

Prior Distribution. During training, we assume that reference distributions of future hand position $Q(\mathcal{M}|x)$ and interaction hotspots $Q(\mathcal{A}|x)$ are known in prior. These distributions can be derived from manual annotation of 2D fingertips and interaction hotspots, as we will describe in Sec 4.1. A 2D isotropic Gaussian is further applied to the annotated 2D points, leading to the distributions of $Q(\mathcal{M}|x)$ and $Q(\mathcal{A}|x)$. If annotations are not available, we adopt uniform distributions for both $Q(\mathcal{M}|x)$ and $Q(\mathcal{A}|x)$.

Variational Learning. Our proposed model seeks to jointly predict motor attention $\mathcal{M}$, interaction hotspots $\mathcal{A}$, and the action label y. Therefore, we inject the posterior $p(\mathcal{A}, \mathcal{M}|x)$ into $p(y|x)$. We further assume $p(\mathcal{A}, \mathcal{M}|x)$ can be factorized into $p(\mathcal{A}|x)$ and $p(\mathcal{M}|x)$ (see supplementary materials for details). Our model thereby optimizes the resulting latent variable model by maximizing the

Evidence Lower Bound (ELBO), given by[4]

$$\begin{aligned} \log p(y|x) \geq &E_{p(\mathcal{A},\mathcal{M}|x)}[\log p(y|\mathcal{A},\mathcal{M},x)] - log(p(\mathcal{A},\mathcal{M}|x))] \\ = &\sum_{\mathcal{A},\mathcal{M}} \log p(y|\mathcal{A},\mathcal{M},x) - KL[p(\mathcal{A}|x)||Q(\mathcal{A}|x)] - KL[p(\mathcal{M}|x)||Q(\mathcal{M}|x)]. \end{aligned}$$
(7)

Therefore, the loss function $\mathcal{L}$ is given by

$$\mathcal{L} = - \sum_{\mathcal{A},\mathcal{M}} \log p(y|\mathcal{A},\mathcal{M},x) + KL[p(\mathcal{A}|x)||Q(\mathcal{A}|x)] + KL[p(\mathcal{M}|x)||Q(\mathcal{M}|x)].$$
(8)

The first term in the loss function is the cross entropy loss for action anticipation. The last two terms use KL-Divergence to align the predicted distributions of motor attention $p(\mathcal{M}|x)$ and interaction hotspots $p(\mathcal{A}|x)$ to their reference distributions ($Q(\mathcal{M}|x)$ and $Q(\mathcal{A}|x)$). To make the training practical, we draw a single sample for each input within a mini-batch similar to [26,37]. Multiple samples of the same input will be drawn at different iterations.

Approximate Inference. At inference time, our model could have drawn many samples of motor attention $\tilde{\mathcal{M}}$ and interaction hotspots $\tilde{\mathcal{A}}$ for the anticipation. However, the sampling and averaging is computationally expensive. We choose to feed deterministic $\mathcal{M}$ and $\mathcal{A}$ into Eq. 5 and Eq. 6 at inference time. Note that F_A and F_P are convex, since they are composed of linear mapping function and softmax function. By Jensen's inequality, we have

$$E[F_A(\tilde{\mathcal{M}},x)] \geq F_A(E[\tilde{\mathcal{M}}],x) = F_A(\mathcal{M},x),$$
(9)

$$E[F_P(\tilde{\mathcal{A}},\tilde{\mathcal{M}},x)] \geq F_P(E[\tilde{\mathcal{A}}],E[\tilde{\mathcal{M}}],x) = F_P(\mathcal{A},\mathcal{M},x)$$
(10)

Therefore, such approximation provides a valid lower bound of $E[F_P(\tilde{\mathcal{A}},\tilde{\mathcal{M}},x)]$ and $E[F_A(\tilde{\mathcal{M}},x)]$, and serves as a shortcut to avoid sampling during testing.

3.6   Network Architecture

We consider two different backbone networks for our model, including lightweight I3D-Res50 network [4,60] pre-trained on Kinetics and heavy CSN-152 [54] network pre-trained on IG-65M [17]. We use I3D-Res50 for our ablation study on EGTEA and EPIC-Kitchens, and report results using CSN-152 backbone when competing on the EPIC-Kitchens dataset. Both networks have five convolutional blocks. The motor attention module, the interaction hotspots module and the recognition module are attached to the 2nd, the 3rd and the 5th block, respectively. We use 3D max pooling to match the size of attention map to the size of the feature map in Eq. 5 and Eq. 6. For training, our model takes an input of 32 frames (every other frame from a 64-frame chunk) with a resolution of 224×224. For inference, our model samples 30 clips from a video (3 along width of frame and 10 in time). Each clip has 32 frames with a resolution of 256×256. We average the scores of all sampled clips for video level prediction. Other implementation details will be discussed in the experiments.

---
[4] See supplementary material for the derivation.

4   Experiments

We now present our experiments and results. We briefly introduce our implementation details and describe the datasets and annotations. Moreover, we present our results on EPIC-Kitchens action anticipation challenge, followed by ablation studies that further evaluate our model on interaction hotspot estimation and motor attention prediction. Finally, we provide a discussion of our method.

Implementation Details. Our model is trained using SGD with momentum 0.9 and batch size 64 on 4 GPUs. The initial learning rate is 2.5e-4 with cosine decay. We set weight decay to 1e-4 and enable batch norm [24]. We downsample all frames to 320x256 (24fps) for EGTEA, and 512x288 (30fps) for EPIC-Kitchens. We apply several data augmentation techniques, including random flipping, rotation, cropping and color jittering to avoid overfitting.

4.1   Datasets and Annotations

Datasets. We make use of two FPV datasets: EGTEA Gaze+ [32,33] and Epic-Kitchens [6]. EGTEA comes with $10,321$ action instances from $19/53/106$ verb/noun/action classes. We report results on the first split of the dataset. EPIC-Kitchens contains $39,596$ instances from 125 verbs and 352 nouns. We follow [15] to split the public training set into training and validation sets with 2513 action classes. We conduct ablation studies on this train/val split, and present the action anticipation results on the testing sets. We set the anticipation time as 0.5 seconds for EGTEA and 1 second [6] for EPIC-Kitchens.

Annotations. Our model requires supervisory signals of interaction hotspots and hand trajectories during training. We provide extra annotations for both EGTEA and EPIC-Kitchens datasets. These annotations will be made publicly available. Specifically, we manually annotated interaction hotspots as 2D points on the last observable frames for all instances on EGTEA and a subset of instances on EPIC-Kitchens. This is because many noun labels in Epic-Kitchens have very few instances, hence we focus on interaction hotspots of action instances that include many-shot nouns [6] in the training set.

Moreover, we explore different approaches to generate the pseudo ground truth of future hand trajectories. On EGTEA, we trained a hand segmentation model ([35] using hand masks from the dataset). The motor attention was approximated by segmenting hands at every frame and tracking the fingertip closest to an active object. To mitigate ego-motion, we used optical flow and RANSAC to compute a homography transform, and project the motor attention to the last observable frame. As EPIC-Kitchens does not provide hand masks, we instead annotated the fingertip closest to an interaction hotspots on the last observable frame. A linear interpolation of 2D motion between the fingertip and the interaction hotspots was used to approximate the motor attention.

4.2   FPV Action Anticipation on EPIC-Kitchens

We highlight our results for FPV action anticipation on EPIC-Kitchens dataset.

Table 1. Action anticipation results on Epic-Kitchens. Ours+Obj model outperforms state-of-the-art by a notable margin. See discussions of Ours+Obj in Sec. 4.2.

	Method	Top1/Top5 Accuracy		
		Verb	Noun	Action
s1	2SCNN [6]	29.76 / 76.03	15.15 / 38.65	4.32 / 15.21
	TSN [6]	31.81 / 76.56	16.22 / 42.15	6.00 / 18.21
	TSN+MCE [14]	27.92 / 73.59	16.09 / 39.32	10.76 / 25.28
	Trans R(2+1)D [38]	30.74 / 76.21	16.47 / 42.72	9.74 / 25.44
	RULSTM [15]	33.04 / **79.55**	22.78 / 50.95	14.39 / 33.73
	Ours	34.99 / 77.05	20.86 / 46.45	14.04 / 31.29
	Ours+Obj	**36.25** / 79.15	**23.83** / **51.98**	**15.42** / **34.29**
s2	2SCNN [6]	25.23 / 68.66	9.97 / 27.38	2.29 / 9.35
	TSN [6]	25.30 / 68.32	10.41 / 29.50	2.39 / 9.63
	TSN+MCE [14]	21.27 / 63.66	9.90 / 25.50	5.57 / 25.28
	Trans R(2+1)D [38]	28.37 / 69.96	12.43 / 32.20	7.24 / 19.29
	RULSTM [15]	27.01 / 69.55	15.19 / 34.38	8.16 / 21.20
	Ours	28.27 / 70.67	14.07 / 34.35	8.64 / 22.91
	Ours+Obj	**29.87** / **71.77**	**16.80** / **38.96**	**9.94** / **23.69**

Experiment Setup. To compete for EPIC-Kitchens anticipation challenge, we used the backbone network CSN152. We trained our model on the public training set and report results using top-1/5 accuracy as in [6].

Results. Table 1 compares our results to latest methods on EPIC-Kitchens. Our model outperforms strong baselines (TSN and 2SCNN) reported in [6] by a large margin. Compared to previous best results from RULSTM [15], our model archives +2%/-1.9%/-0.3% for verb/noun/action on seen set, and +1.3%/-1.1%/+0.6% on unseen set of EPIC-Kitchens. Our results are better for verb, worse for noun and comparable or better for actions. Notably, RULSTM requires object boxes & optical flow for training and object features & optical flow for testing. In contrast, our method uses hand trajectories and interaction hotspots for training and needs *only RGB frames* for testing.

To further improve the performance, we fuse the object stream from RUL-STM with our model (Ours+Obj). Compared to RULSTM, Ours+Obj has a performance gain of +3.2%/+2.9% for verb, +1.1%/+1.6% for noun, and +1.0%/+1.8% for action (seen/unseen). It is worthy pointing out that RUL-STM benefits from an extra flow network, while ours+Obj model takes additional supervisory signals of hands and hotspots. Note that our performance boost does not simply come from those extra annotations. In a subsequent ablation study, we have shown that simply training with these extra annotations has minor improvement, when used without our proposed probabilistic deep model.

We note that it is not possible to make a direct apples-to-apples comparison between our model and RULSTM [15], as the two models used vastly different training signals. We refer readers to the supplementary materials for a detailed experiment setup comparison. In terms of performance, our model is comparable to RULSTM without using any side information for inference. When using addi-

Table 2. Ablation study for action anticipation. We compare our model with backbone I3D network, and further analyze the role of motor attention prediction, interaction hotspots estimation, and stochastic units in joint modelling. See discussions in Sec. 4.3.

Method	EGTEA			Epic-Kitchens		
	Top1 Accuracy / Mean Cls Accuracy			Top1 Accuracy / Top5 Accuracy		
	Verb	Noun	Action	Verb	Noun	Action
I3D-Res50	48.01/31.25	42.11/30.01	34.82/23.20	30.06/76.86	16.07/41.67	9.60/24.29
JointDet	48.58/32.21	43.95/31.26	35.69/23.59	30.16/**76.86**	16.25/41.71	9.76/24.40
Hotspots Only	47.95/31.94	44.02/32.53	35.50/23.82	30.21/75.93	16.57/42.28	9.66/24.33
Motor Only	**49.35**/32.34	**45.69/33.93**	36.49/25.13	30.63/76.69	17.28/42.56	10.21/25.32
Ours	48.96/**32.48**	45.50/32.73	**36.60/25.30**	**30.65**/76.53	**17.40/42.60**	**10.38/25.48**

Table 3. Ablation study for interaction hotspots estimation. Jointly modeling motor attention with stochastic units can greatly benefit the performance of interaction hotspots estimation. ($\uparrow/\downarrow$ indicates higher/lower is better) See discussions in Sec. 4.3.

Method	EGTEA				Epic-Kitchens			
	Prec $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	KLD $\downarrow$	Prec $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	KLD $\downarrow$
I3DHeatmap	12.82	37.53	19.11	2.66	17.20	77.39	28.15	3.07
JointDet	16.11	41.82	23.26	1.84	17.32	85.79	28.83	2.21
Ours	**17.43**	**48.81**	**25.69**	**1.62**	**17.86**	**86.59**	**29.60**	**1.99**

tional object stream during inference as in RULSTM, our model outperforms RULSTM by a relative improvement of **7%/22%** on seen/unseen set. More importantly, our model also provides the additional capabilities of predicting future hand trajectories and estimating interaction hotspots.

4.3   Ablation Study

We present ablation studies of our model. We introduce our experiment setup, evaluate each component of our model, and then contrast our method to a series of baselines on motor attention prediction and interaction hotspot estimation **Experiment Setup**. For all of our ablation studies, we adopt the lightweight I3D-Res50 [60] as backbone network to reduce computational cost. Our model is evaluated for action anticipation, motor attention prediction and interaction hotspots estimation across EGTEA (using split1) and EPIC-Kitchens (using the train/val split from [15]). Specifically, we consider the following metrics.

- **Action Anticipation**. We report Top1/Mean Class accuracy on EGTEA as in [34] and Top1/Top5 accuracy as on EPIC-Kitchens following [15].
- **Interaction Hotspots Estimation**. We report F1 score as in [32] and KL-Divergence (KLD) as in [39] using a downsampled heatmap (32x) at the last observable frame.
- **Motor Attention Prediction**. We report the average and final displacement errors between the most confident location on a predicted attention map and the ground-truth hand points, similar to previous work on trajectory prediction [3]. Note that the motor attention maps is downsampled by a factor of 32/8 in space/time. Hence, we report displacement errors normalized in spatial and temporal dimension.

Benefits of Joint Modeling. As a starting point, we compare our model with a backbone I3D-Res50 model. We present the results of action anticipation in Table 2. In comparison to I3D-Res50, our model improves noun and action prediction by +3.4%/1.8% on EGTEA and +1.3%/0.8% on EPIC-Kitchens. Moreover, we show that our model improves the performance of interaction hotspots estimation. We consider the baseline I3D model that only estimates interaction region with interaction hotspots module as I3DHeatmap. As shown in Table 3, our model improves the F1 score by 6.6%/1.5% on EGTEA/EPIC-Kitchens.

Stochastic Modeling vs. Deterministic Modeling. We further evaluate the benefits of probabilistic modeling of motor attention and interaction hotspots. To this end, we compare our model with a deterministic joint model (*Joint-Det*). JointDet has the same architecture as our model, except for the stochastic units. As shown in Table 2, JointDet slightly improve the I3D baseline for action anticipation (+0.87% on EGTEA and +0.16% on EPIC-Kitchens), yet lags behind our probabilistic model. Specifically, our model outperforms JointDet by 0.91% and 0.62% on EGTEA and EPIC-Kitchens. Moreover, in comparison to JointDet, our model has better performance for interaction hotspots estimation (+2.4%/ + 0.8% in F1 scores on EGTEA/EPIC-Kitchens). These results suggest that simply training with extra annotations might fail to capture the uncertainty of visual anticipation. In contrast, our design choice of probabilistic modeling can effectively deal with those uncertainty, therefore helps to improve the performance of joint modeling.

Motor Attention vs. Interaction Hotspots. Futhermore, we evaluate the contributions of motor attention and interaction hotspots for FPV action anticipation. We consider two baseline models in Table 3: I3D model equipped with only motor attention module (*Motor Only*), and I3D model equipped with only interaction hotspots module (*Hotspots Only*). Both models underperform the full model across the two datasets, yet the gap between *Motor Only* and the full model is smaller. These results suggest that both components contribute to the performance boost of action anticipation, yet the modeling of motor attention weights more than the modeling of interaction hotspots.

Interaction Hotspots Estimation. We present additional results on interaction hotspots estimation. We compare our results to the following baselines.

- **Center Prior** represents a Gaussian Distribution at the center of the image.
- **Grad-Cam** uses the same I3D backbone network as our model, and produces a saliency map via Grad-Cam [48].
- **EgoGaze** considers possible gaze position as salient region of a given image. This model is trained on eye fixation annotation from EGTEA-Gaze+ [23]. The assumption is that the person is likely to look at the interaction hotspots.
- **DSS Saliency** predicts salient region during human object interaction. This model is trained on pixel-level saliency annotation from [22].
- **EgoHotspots** is the latest work [39] for estimating interaction hotspots.

Our results are shown in Table 4. Our model outperforms the best baselines (EgoGaze and EgoHotspots) by 5.4% on EGTEA and 3.6% on EPIC-Kitchens

Table 4. Interaction hotspots estimation results on EGTEA and EPIC-Kitchens. Our model outperforms a set of strong baselines. ($\uparrow$/$\downarrow$ indicates higher/lower is better)

Method	EGTEA				Epic-Kitchens			
	Prec$\uparrow$	Recall$\uparrow$	F1$\uparrow$	KLD$\downarrow$	Prec$\uparrow$	Recall$\uparrow$	F1$\uparrow$	KLD$\downarrow$
Center Prior	10.87	17.65	13.45	10.64	11.66	16.97	13.82	10.27
Grad-Cam [48]	9.98	22.13	13.76	8.73	10.85	20.01	14.07	8.06
DSS [22]	9.02	39.49	14.69	6.12	12.03	33.75	17.74	5.21
EgoGaze [23]	15.02	31.34	20.31	3.20	11.30	27.65	16.05	3.37
EgoHotspots [39]	16.51	24.07	19.59	3.36	**22.26**	31.37	26.04	2.84
Ours	**17.43**	**48.81**	**25.69**	**1.62**	17.86	**86.5**	**29.6**	**1.99**

Table 5. Motor attention prediction results on EGTEA. Our model compares favourably to strong baselines. ($\uparrow$/$\downarrow$ indicates higher/lower is better)

Method	Avg. Disp. Error $\downarrow$	Final Disp. Error $\downarrow$
Kalman Filter	0.32	0.48
GPR	0.29	0.37
LSTM	**0.22**	**0.35**
Ours	0.23	0.36

in F1 scores. These results suggest that our proposed joint model can effectively identify future interaction region. Another observation is that our model performs better on EPIC-Kitchens than EGTEA. This is probably due to the larger number of available training samples.

Motor Attention Prediction. We report our results on motor attention prediction. We consider the following baselines and only report results on EGTEA, as the future hand position on EPIC-Kitchens is not accurate (see Sec. 4.1).

- **Kalman Filter** describes the hand trajectory prediction problem with state-space model, and assumes linear acceleration during update step.
- **Gaussian Process Regression (GPR)** iteratively predicts the future hand position using Gaussian Process Regression.
- **LSTM** adopts a vanilla LSTM network for trajectory forecasting. We use the implementation from [3].

The results are presented in Table 5. Our model outperforms Kalman filter and GPR, yet is slightly worse than LSTM model (+0.01 in both errors). Note that all baseline methods need the coordinate of the first observed hand for prediction. This simplifies trajectory prediction into a less challenging regression problem. In contrast, our model does not need hand coordinates for inference. A model that relies on the observation of hand positions will encounter failure cases when the hand has not been observed, while our model is still capable of "imagining" the possible hand trajectory. See "Operate Microwave" and "Wash Coffee Cup" in Fig. 3 for example results from our model.

Visualization of Motor Attention and Interaction Hotspots. Finally, we visualize the predicted motor attention, interaction hotspots, and action labels from our model in Fig. 3. The predicted motor attention almost always attends to the predicted objects and corresponding interaction hotspots. Hence, our model can address challenging cases where next-active objects are ambiguous. Take the
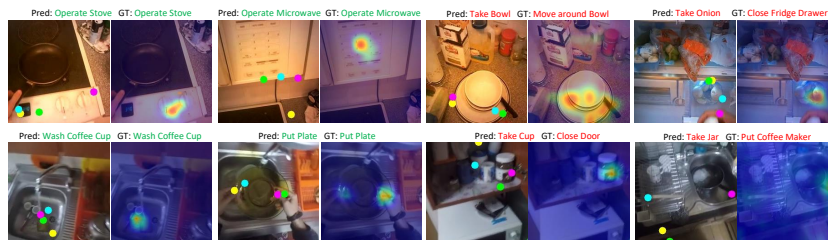
Fig. 3. Visualization of motor attention (left image), interaction hotspots (right image), and action labels (captions above the images) on sample frames from EGTEA (first row) and EPIC-Kitchens (second row). Both successful (green label) and failure cases (red label) are shown. Future hands position are predicted at every 8 frames and plotted on the last observable frame with the order of yellow, green, cyan, and magenta.

first example of "Operate Stove" in Fig. 3. Our model successfully predicted the future objects and estimated the interaction hotspots as the stove control knob.

4.4   Remarks and Discussion

We must also point out that our method has certain limitations, which point to exciting future research directions. For example, our model requires additional annotations for training, which might bring scalability issues when analyzing other datasets. These dense annotations can indeed be approximated using sparsely annotated frames as discussed in Sec. 4.1. We speculate that more advanced hand tracking and object segmentation models can be explored to generating the pseudo ground truth of motor attention and interaction hotspots. Moreover, our model shares a similar conundrum faced by previous work on anticipation. Our model is likely to fail when future active objects are not observed. See "Close Fridge Drawer" and "Put Coffee Maker" in Fig. 3. We conjecture that these cases requires incorporating logical reasoning into learning based methods—an active research topic in our community.

5   Conclusions

We presented the first deep model that jointly predicts motor attention, interaction hotspots, and future action labels in FPV. Importantly, we demonstrated that motor attention plays an important role in forecasting human-object interactions. Another key insight is that characterizing motor attention and interaction hotspots as probabilistic variables can account for the stochastic pattern of human intentional movement. We believe that our model provides a solid step towards the challenging problem of visual anticipation.

References

1. Aglioti, S.M., Cesari, P., Romani, M., Urgesi, C.: Action anticipation and motor resonance in elite basketball players. Nature neuroscience (2008)
2. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: ICCV (2019)
3. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: Human trajectory prediction in crowded spaces. In: CVPR (2016)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
5. Chen, C.Y., Grauman, K.: Subjects and their objects: Localizing interactees for a person-centric view of importance. International Journal of Computer Vision **126**(2-4), 292–313 (2018)
6. 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: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV. pp. 753–771. Springer International Publishing (2018)
7. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV. pp. 284–298. Springer Berlin Heidelberg (2012)
8. Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G.: Understanding motor events: a neurophysiological study. Experimental brain research (1992)
9. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: CVPR (2018)
10. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV (2011)
11. Felsen, P., Agrawal, P., Malik, J.: What will happen next? forecasting player moves in sports videos. In: ICCV (2017)
12. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV (2015)
13. Furnari, A., Battiato, S., Grauman, K., Farinella, G.M.: Next-active-object prediction from egocentric videos. VCIP (2017)
14. Furnari, A., Battiato, S., Maria Farinella, G.: Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In: Leal-Taixé, L., Roth, S. (eds.) ECCV Workshops. pp. 389–405. Springer International Publishing (2018)
15. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In: ICCV (2019)
16. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. In: BMVC (2017)
17. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: CVPR (2019)
18. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR (2011)
19. Gui, L.Y., Wang, Y.X., Liang, X., Moura, J.M.: Adversarial geometry-aware human motion prediction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV. pp. 823–842. Springer International Publishing (2018)
20. Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., Rizzolatti, G.: Activation of human primary motor cortex during action observation: a neuromagnetic study. Proceedings of the National Academy of Sciences (1998)

21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
22. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: CVPR (2017)
23. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV. pp. 789–804. Springer International Publishing (2018)
24. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
25. James, W., Burkhardt, F., Bowers, F., Skrupskelis, I.K.: The principles of psychology, vol. 1. Macmillan London (1890)
26. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017)
27. Kataoka, H., Miyashita, Y., Hayashi, M., Iwata, K., Satoh, Y.: Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature. In: BMVC (2016)
28. Ke, Q., Fritz, M., Schiele, B.: Time-conditioned action anticipation in one shot. In: CVPR (2019)
29. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV. pp. 201–214. Springer Berlin Heidelberg (2012)
30. Kong, Y., Fu, Y.: Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230 (2018)
31. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE transactions on pattern analysis and machine intelligence **38**(1), 14–29 (2015)
32. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV. pp. 639–655. Springer International Publishing (2018)
33. Li, Y., Liu, M., Rehg, J.M.: In the eye of the beholder: Gaze and actions in first person video. arXiv preprint arXiv:2006.00626 (2020)
34. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: CVPR (2015)
35. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
36. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR (2016)
37. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: ICLR (2017)
38. Miech, A., Laptev, I., Sivic, J., Wang, H., Torresani, L., Tran, D.: Leveraging the present to anticipate the future in videos. In: CVPR Workshops (2019)
39. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: ICCV (2019)
40. Pavlovic, V., Rehg, J.M., MacCormick, J.: Learning switching linear models of human motion. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) NeurIPS. pp. 981–987. MIT Press (2001)
41. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012)
42. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact CNN for indexing egocentric videos. In: WACV (2016)
43. Rhinehart, N., Kitani, K.M.: Learning action maps of large environments via first-person vision. In: CVPR (2016)

44. Rhinehart, N., Kitani, K.M.: First-person activity forecasting with online inverse reinforcement learning. In: ICCV (2017)
45. Rushworth, M., Johansen-Berg, H., Göbel, S.M., Devlin, J.: The left parietal and premotor cortices: motor attention and selection. Neuroimage **20**, S89–S100 (2003)
46. Ryoo, M.S., Rothrock, B., Matthies, L.: Pooled motion features for first-person videos. In: CVPR (2015)
47. Ryoo, M., Fuchs, T.J., Xia, L., Aggarwal, J.K., Matthies, L.: Robot-centric activity prediction from first-person videos: What will they do to me? In: HRI (2015)
48. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
49. Shen, Y., Ni, B., Li, Z., Zhuang, N.: Egocentric activity prediction via event modulated attention. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Proceedings of the European Conference on Computer Vision (ECCV). pp. 202–217. Springer International Publishing (2018)
50. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
51. Soo Park, H., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: CVPR (2016)
52. Soran, B., Farhadi, A., Shapiro, L.: Generating notifications for missing actions: Don't forget to turn the lights off! In: ICCV (2015)
53. Thermos, S., Papadopoulos, G.T., Daras, P., Potamianos, G.: Deep affordance-grounded sensorimotor object recognition. In: CVPR (2017)
54. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: ICCV (2019)
55. Urtasun, R., Fleet, D.J., Geiger, A., Popović, J., Darrell, T.J., Lawrence, N.D.: Topologically-constrained latent variable models. In: ICML (2008)
56. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR (2016)
57. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV (2017)
58. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE transactions on pattern analysis and machine intelligence **30**(2), 283–298 (2007)
59. Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: CVPR (2017)
60. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
61. Wei, P., Xie, D., Zheng, N., Zhu, S.C.: Inferring human attention by learning latent intentions. In: IJCAI (2017)
62. Yagi, T., Mangalam, K., Yonetani, R., Sato, Y.: Future person localization in first-person videos. In: CVPR (2018)
63. Zhou, Y., Ni, B., Hong, R., Yang, X., Tian, Q.: Cascaded interactional targeting network for egocentric video analysis. In: CVPR (2016)