Forecasting Human Object Interaction: Joint Prediction of Motor Attention and Egocentric Activity

Miao Liu¹,³, Siyu Tang³,⁴, Yin Li², and James M. Rehg¹

¹Georgia Institute of Technology, Atlanta, United States
²University of Wisconsin-Madison, Madison, United States
³Max Planck Institute for Intelligent Systems, Tübingen, Germany
⁴University of Tübingen, Tübingen, Germany

Abstract

We address the challenging task of anticipating human-object interaction in first person videos. Most existing methods ignore how the camera wearer interacts with the objects, or simply consider body motion as a separate modality. In contrast, we observe that the international hand movement reveals critical information about the future activity. Motivated by this, we adopt intentional hand movement as a future 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 latent variables in our deep model. The predicted motor attention is further used to characterise 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. At the time of submission, our method is ranked first on unseen test set during EPIC-Kitchens Action Anticipation Challenge Phase 2.

1. Introduction

Action anticipation remains a key challenge in computer vision. However, we humans can easily address the task of “looking into the near future”. Consider the example in Fig. 1, given a video clip shortly before the start of an action, we can easily predict what is going to happen next, e.g., the person will take the salt canister. More interestingly, without seeing future frames, we can vividly imagine the details of how the person will perform the action, e.g.,

Figure 1. What is the most likely action? 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) from the last observable time step to action starting point. Visualizations of hand trajectory are forecasted to the last observable frame. (Best view in color)

the trajectory of the hand when reaching the canister and the part of the canister to grasp. In fact, our remarkable ability to forecast other individuals’ actions critically depends on our perception and interpretation of their body motion.

The investigation of this anticipatory mechanism can be traced back to 19th century. William James argued that future expectations are intrinsically related to purposive body movements [24]. Additional evidence for 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, which is the same brain region 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]. The compelling explanation stated in [46] 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 findings in cognitive neuroscience, the intentional body motion is largely ignored by existing action anticipation literatures [57, 11, 16, 26, 13, 15, 37, 27]. In this work, we focus on the problem of forecasting human-object interactions in First Person Vision (FPV). 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, 58]. Our work shares the same motivation of future forecasting, yet we are the first to incorporate motor attention for FPV action anticipation.

To this end, we propose a novel neural network that predicts “motor attention”—the future trajectory of the hands, as an anticipatory representation of actions. Based on motor attention, our model further recognizes the type of the interaction and localizes the contact point of the interaction, i.e., interaction hotspots [38]. Importantly, we characterize motor attention and interaction hotspots as latent variables modeled by stochastic units in the network. These units naturally deal with the uncertainty of future hand motion and human object interaction, and produce attention maps that highlight discriminative spatial-temporal features for interaction anticipation.

During inference, our model takes video clips shortly before the interaction begins as inputs, and predicts the future hand trajectory, the interaction category, and the location of the hotspots. During training, our model assumes that these outputs are available as the supervisory signals. We evaluate our model on two major FPV benchmarks: EGTEA Gaze+ and EPIC-Kitchens. We conduct detailed ablation studies to show the effectiveness of our model. Importantly, our results of action anticipation outperform the state-of-the-art by a significant margin on both datasets. At the time of submission, our results ranked the first on unseen test and the second on seen test set of EPIC-Kitchens. Moving beyond actions, we evaluate our model for hand trajectory prediction and interaction hotspot detection. Our model demonstrates strong results for both tasks. We believe that our model provides a solid step forward towards 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, and some other anticipation tasks. We also review some recent efforts on visual annoyance.

FPV Action Anticipation. Action anticipation indicates the task of predicting an action before it happens. We refer the readers to a recent survey [29] for a detailed discussion on the difference between action recognition and anticipation. The action recognition in egocentric videos has been studied extensively [48, 42, 10, 62, 35, 32, 31, 41], while fewer works target at egocentric action anticipation. Shen et al. [50] investigated how different egocentric modalities affect the action anticipation performance. Soran et al. [53] adopts Hidden Markov Model to compute the transition probability among a sequences of actions. A similar idea was explored in [37]. Furnari et al. [13] introduced the task of predicting next-active objects. They used object trajectories to discriminate active objects from passive objects, and thereby predict the activated temporal segment. 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. [27] proposed a time-conditioned skip connection operation to extract relevant information from observable video sequence. In contrast to our proposed method, those 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 investigations seek to address the task of action anticipation in third person vision. Pei et al. [40] proposed to predict the intent of an activity by parsing video events based on a Stochastic Context Sensitive Grammar. Kris et al. [28] combined semantic scene labeling with a Markov Decision Process to forecast the behavior and trajectory of the subject. Vondrick et al. [57] proposed to predict the future video representation from large scale unlabeled video data. Felson et al. [11] developed a generic framework to forecast future events in sports videos. Gao et al. [16] proposed a Reinforced Encoder-Decoder network to summarize past frames representation and produce a hypothesis of future action. Kataoka et al. [26] introduced a subtle motion descriptor to identify the difference between on-going action and transitional action, and therefore facilitate future anticipation. Our work shares the same objective of future forecasting, however our goal is to exploit the abundant visual cues from egocentric videos for action anticipation.

Other Prediction Tasks. Previous works considered future anticipation under various other settings. Rhinehart et al. [44] proposed to an online learning algorithm to forecast the first-person trajectory. Park et al. [52] proposed a fully convolutional neural network to infer possible human tra-
Figure 2. Overview of our model. We adopt 3D ResNet-50 as our backbone network (a). The motor attention module (b) utilizes stochastic units to generate sampled motor attention, which is further used to guide interaction hotspots estimation in module (c). Module (c) further generates sampled interaction hotspots with a similar stochastic units as in module (b). Both sampled motor attention and sampled interaction hotspots 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 $\odot$ represents element-wise multiplication.

Table: Visual Affordance Table

| Visual Affordance | Description |
|-------------------|-------------|
| Visual Affordance | Attracts growing interest in computer vision, since it incorporates important knowledge for scene understanding. Human-object interaction recognition and action anticipation are important for understanding human-object interactions. Previous works only targeted synthetic datasets or less challenging FPV datasets. Most importantly, these prior works did not target the analysis of body movement in the context of action anticipation. |

3. Method

We adopt the same egocentric action anticipation setting defined in [6]. Assume the action segment starts at time step $\tau_s$. Our goal is to predict the activity labels by observing a $\tau_o$ seconds video clip that precedes the action by $\tau_a$ seconds. Formally, we denote $\tau_s$ as the “anticipation time” and $\tau_o$ as the “observation time”. Given the observable video segment $x: \{\tau_s - (\tau_a + \tau_o), \tau_s - \tau_a\}$, we seek to predict the action label $y$ that begins at $\tau_s$. In addition, our model also outputs the interaction hotspots $A$ at time step $\tau_a - \tau_o$ (the last observable frame), and motor attention $M$ from time step $\tau_s - \tau_a$ to time step $\tau_o$. We refer readers to Fig. 1 for a visual illustration of our problem setting.

Fig 2 presents an overview of our model. We denote the backbone 3D convolutional network (a) as $\phi$, and network features from the $i$th convolutional block as $\phi_i(x)$. We assume prior distribution of future hand position $Q(M|x)$ and interaction hotspots $Q(A|x)$ are available for training. However, the patterns of future hand movements and human-object interactions are highly uncertain. Even given the full video, the annotation of hand trajectory and interac-
tion hotspots can be inaccurate. To address this challenge, we model motor attention and interaction hotspots as probabilistic variables to account for their uncertainty. Specifically, the motor attention module (b) predicts motor attention $M$ and uses stochastic units to sample from $M$. The sampled motor attention $\tilde{M}$ serves an indicator of important spatial-temporal features for interaction hotspots estimation. The interaction hotspots module (c) produces interaction hotspots distribution $A$ and its sampled version $\tilde{A}$. The anticipation module (d) further uses both $M$ and $A$ to aggregate network features and predicts the action label $y$.

### 3.1. Deep Latent Variable Model

For our joint model, we consider motor attention $M$ and interaction hotspots $A$ as probabilistic variables, and incorporate two posteriors into our joint model:

$$p(y|x) = \int_M p(y|M, x)p(M|x)dM. \quad (1)$$

$$p(M|x) \text{ and } p(A|M, x) \text{ seem intractable at first glance. Fortunately, variational inference comes to the rescue. In the following sections, we show how the above two posteriors can be parameterized by the network. Our model has three major components:}

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

- **Interaction Hotspots Module** targets at $p(A|M, x)$. Our model uses the interaction hotspots estimation function $F_A$ to estimate the interaction hotspots $A$ based on the network feature $\phi_3(x)$ and sampled motor attention $\tilde{M}$. $A$ is represented as a 2D attention map of size $H_a \times W_a$. A further normalization constrained that $\sum_{w, h} A(w, h) = 1$.

- **Anticipation Module** utilizes motor attention and interaction hotspots for action anticipation. Specifically, sampled motor attention $\tilde{M}$ and sampled interaction hotspots $\tilde{A}$ are further used to aggregate feature $\phi_5(x)$ via weighted pooling. The action anticipation function $F_P$ further maps the aggregated feature 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 $W_M$ and a softmax function. $W_M$ maps 4D network feature $\phi_5(x)$ into a 3D tensor. As shown in Fig 2 module (b), $W_M$ is implemented by 3D convolution. The softmax function further constrains this tensor to be a probabilistic distribution on each temporal slice. We denote this tensor as $\psi$, which is given by $\psi = softmax(W_M(\phi_2(x)))$. Therefore, $p(M|x)$ is given by

$$M_{m,n,t} = \frac{\psi_{m,n,t}}{\sum_{m,n} \psi_{m,n,t}}, \quad (2)$$

where $\psi_{m,n,t}$ represents the expectation of motor attention at spatial position $(m, n)$ and time step $t$.

**Stochastic Modeling.** Modeling motor attention in the context of forecasting human-object interaction requires a mechanism for addressing the stochastic nature of 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 reparameterization trick introduced in [25, 36] to design a differentiable sampling mechanism:

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

where $G$ refers to the Gumbel distribution that can be used to sample from discrete distribution. Gumbel-Softmax can further generate a “soft” sample that allows the direct back-propagation of gradients to $\psi$. $\theta$ is the temperature parameter, which controls the “sharpness” of the distribution. We set $\theta = 2$ for all of our experiments.

### 3.3. Interaction Hotspots Estimation

Motor attention is used to guide interaction hotspots estimation. $p(A|M, x)$ uses sampled motor attention $\tilde{M}$, instead of $M$, as a spatial-temporal saliency map to highlight feature representation $\phi_3(x)$. The interaction hotspots estimation function $F_A$ is composed of a linear function $W_A$ and a softmax function:

$$p(A|M, x) = softmax \left(W_A(\tilde{M} \circ \phi_3(x))\right) \quad (4)$$

where $\circ$ denotes weighted average pooling on individual channel. Given $p(A|M, x)$, we can easily model the conditional probability $p(A|x)$ by

$$p(A|x) = \int_M p(A|M, x)p(M|x)dM. \quad (5)$$

$p(M|x)$ is given by Eq.2. The resulting $A$ is a 2D probabilistic distribution. We then adopt a similar sampling mechanism as in Eq.3. Hence, we have:

$$\tilde{A}_{m,n} \sim \frac{\exp((\log \pi_{m,n} + G_{m,n})/\theta)}{\sum_{m,n} \exp((\log \pi_{m,n} + G_{m,n})/\theta)}. \quad (6)$$

where $\pi_{m,n}$ is given by Eq. 5. It represents the expectation of interaction hotspots at position $(m, n)$.

### 3.4. Egocentric Action Prediction

In previous sections, we have introduced the modeling of $p(M|x)$ and $p(A|M, x)$. The only missing piece is
where $W_P$ is a linear function that predicts action labels.  $\hat{A}$ represents sampled interaction hotspots on the last observable frame, therefore the summation over Eq. 7 is only enforced on the last temporal slice of $M$.

3.5. Training and Inference

**Variational Learning.** Our proposed model seeks to jointly predict motor attention $M$, interaction hotspots $A$ and the action label $y$. Therefore, we deliberately inject posterior $p(A,M|x)$ into $p(y|x)$ and optimize the resulting latent variable model by maximizing the Evidence Lower Bound (ELBO):

$$\log p(y|x) \geq -\mathcal{L} = E_{p(A,M|x)}[\log p(y|A,M,x)] - \log p(A,M|x)]$$

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

*see supplementary material for proof

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

$$- KL[p(M|x)||Q(M|x)]$$

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

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

$$+ KL[p(M|x)||Q(M|x)]$$

Comparing Eq. 7 with Eq. 9, we can conclude that the first term on the right hand side of Eq. 9 is cross entropy loss using features aggregated by motor attention $M$ and interaction hotspots $A$. The remaining two terms enforce the model to match predicted motor attention and interaction hotspots with corresponding prior distributions. Note that we set $Q(M|x)$ and $Q(A|x)$ as uniform distributions, when the annotation is not available for certain samples.

**Approximate Inference.** At inference time, our model should draw many motor attention samples $\hat{M}$ and interaction hotspots samples $\hat{A}$. For high dimensional video inputs $x$, this process can be computationally expensive. We choose to directly feed deterministic $M$ and $\hat{A}$ into Eq. 6 and Eq. 7. As introduced in the previous section, both $F_{\hat{A}}$ and $F_P$ are convex, since they are composed of linear mapping function and softmax function. By Jensen’s Inequality:

$$E[F_{A}(\hat{M}, x)] \geq F_{A}(E[\hat{M}], x) = F_{A}(M, x)$$

$$E[F_{P}(\hat{A}, \hat{M}, x)] \geq F_{P}(E[\hat{A}], E[\hat{M}]x) = F_{P}(A, M, x)$$

Therefore, such approximation provides valid lower bound of $E[F_{A}(\hat{M}, x)]$ and $E[F_{P}(A, M, x)]$ and serves as a shortcut to avoid dense sampling.

3.6. Implementation Details

Our model uses the I3D ResNet50 network with pre-trained weights from [4] as the backbone. We downsample all frames to $320 \times 256$ with 24 fps for the EGTEA dataset, and $512 \times 288$ with 30 fps for the EPIC-Kitchens dataset. We apply several data augmentation techniques, including random flipping, rotation, cropping and color jittering to avoid overfitting. Our model takes 32 consecutive frames (subsampled by 2 in the temporal dimension) as inputs, and all frames are cropped to $224 \times 224$ for training. Our model is trained using SGD with momentum of 0.9 and a batch size of 64 on 4 GPUs. The initial learning rate is 0.00025 with cosine decay. We set weight decay to 1e-4 and also enable batch norm [23]. Our model is implemented in PyTorch and the code will be made publicly available. For testing, our model takes video clips with spatial resolution $256 \times 256$ as inputs, and applies spatial-temporal resampling. We then average the scores of all resampled instances to get the final prediction results.

4. Experiments

4.1. Dataset and Annotation

**Dataset.** We make use of two FPV benchmark datasets: EGTEA Gaze+ [31] and Epic-Kitchens [6]. EGTEA comes with action annotations of 10321 instances from 19 verb classes, 53 verb classes, and 106 action classes. We use the first split (8,299 for training, 2,022 for testing) of the dataset to evaluate the performance of our method. EPIC-Kitchens contains 39,596 instances from 125 verbs, and 352 nouns. We follow [15] to split the public training set (28,472 instances) into training (23,493 instances) and validation (4,979 instances) sets, and define 2513 action classes. We conduct ablation studies on this training/validation split, and present the action anticipation results on the testing sets. For the EGTEA dataset, we set anticipation time as 0.5 seconds. For the EPIC-Kitchens dataset, we set anticipation time as 1 second as defined in the Anticipation Challenge.

**Data Annotation.** Our model requires various supervisory signals during training. We first annotate interaction hotspots on the last observable frames on EGTEA and EPIC-Kitchens datasets. Since many nouns labels in Epic-Kitchens have very few instances, we only provide interaction hotspots annotation to all many-shot nouns (defined in [6]) in the training data. To simplify the problem, we only consider the motor attention of one hand. The EGTEA dataset has hand mask annotation, so we use the future trajectory of fingertip that is the closest to the future active objects to represent motor attention. To mitigate the back-
Motor Attention Prediction

score is a more suitable metric than AUC-J, which is used in binary pixel labeling problems, where each instance only has one label. Interaction hotspots can be considered as a long-tailed distribution problem as in [31] and KL-Divergence (KLD) as in [38]. Estimating motor attention and interaction hotspots can facilitate future research in FPV human-object interaction.

4.2. Evaluation Metrics

We now elaborate on our evaluation metrics for the proposed prediction tasks.

**Action Anticipation:** We report Top1/Mean Class accuracy on EGTEA as in [31] and Top1/Top5 accuracy as in EPIC-Kitchens Action Anticipation Challenge [6].

**Interaction Hotspots Estimation:** We downsample interaction hotspots by a spatial factor of 32 and report F1 score as in [31] and KLD-Divergence (KLD) as in [38]. Estimating interaction hotspots can be considered as a long-tailed binary pixel labeling problem, where each instance only has a small amount of True Positive pixels. We argue that F1 score is a more suitable metric than AUC-J, which is used by some recent affordance detection studies [38].

**Motor Attention Prediction:** We downsample motor attention by a spatial factor of 32 and temporal factor of 8. We consider the pixel position with highest confidence score as the predicted future fingertip position at each time slice. We then use average displacement error and final displacement error similar to previous trajectory prediction studies [3]. We evaluate our model in pixel space, instead of real world coordinates as in [3].

4.3. Ablation Study

We start with an ablation study of our proposed model. Specifically, we assess the role of stochastic units, motor attention and interaction hotspots in our proposed model.

- **Joint Modeling vs. I3D Backbone:** We adopt the I3D [4] model as our backbone network. The I3D model, even if equipped with 3D convolution for temporal knowledge reasoning, performs poorly on action prediction in comparison to action recognition [31, 4]. This is because the features of the video clip preceding the action are not discriminative enough. Our model outperforms I3D network across all anticipation tasks and datasets by a large margin. The results are presented in Table 1. Our model yields major improvements on both datasets. (↑ / ↓ indicates higher/lower is better)

- **Stochastic Modeling vs. Deterministic Modeling:** To better understand how motor attention and interaction hotspots contribute to the performance boost, we keep the same stochastic modeling branch from our model, and report the performance in Table 1 (named Ours-MO). Our-MO only lags behind our full model by a small gap on EPIC-Kitchen, and even works slightly better on EGTEA. This suggests that most of the performance boost comes from the modeling of motor attention. This again supports our claim that motor attention plays an important role in future anticipation. In contrast, interaction hotspots estimation has minor impact on action anticipation. This is because motor attention itself already includes important knowledge of interaction.

- **Motor Attention vs. Interaction Hotspots:** To better understand how motor attention and interaction hotspots contribute to the performance boost, we keep the same motor attention branch and remove the hotspots estimation branch from our model, and report the performance in Table 1 (named Ours-MO). Our-MO only lags behind our full model by a small gap on EPIC-Kitchen, and even works slightly better on EGTEA. This suggests that most of the performance boost comes from the modeling of motor attention. This again supports our claim that motor attention plays an important role in future anticipation. In contrast, interaction hotspots estimation has minor impact on action anticipation. This is because motor attention itself already includes important knowledge of interaction.

- **Stochastic Modeling vs. Deterministic Modeling:** The random nature of human intentional movement incurs major
Table 3. Action anticipation results on Epic-Kitchen test sets. Our proposed model outperforms previous results by a large margin.

| Method   | Verb Top1/Top5 Accuracy | Noun Top1/Top5 Accuracy | Action Top1/Top5 Accuracy |
|----------|-------------------------|-------------------------|---------------------------|
| 2SCNN [6] | 29.76 / 76.03           | 15.15 / 38.65           | 4.32 / 15.21              |
| TSN+MCE [14] | 31.81 / 76.56         | 16.22 / 42.15           | 6.00 / 18.21              |
| Trans R(2+1)D [37] | 27.92 / 73.59     | 16.09 / 39.32           | 10.76 / 25.28             |
| RULSTM [15] | 30.74 / 76.21          | 16.47 / 42.72           | 9.74 / 25.44              |
| Ours     | 33.04 / 79.55          | 22.78 / 50.95           | 14.39 / 33.73             |
| Ours+Obj | 34.99 / 77.05          | 20.86 / 46.45           | 14.04 / 31.29             |
| 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 [37] | 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              |

Table 4. Interaction hotspots estimation results on EGTEA and Epic-Kitchens. Our model outperforms a series of baseline results by a significant margin. (↑ / ↓ indicates higher/lower is better)

| Method      | EGTEA          | Epic-Kitchens       |
|-------------|----------------|---------------------|
| Center Prior | 10.87 / 7.65   | 11.66 / 10.27       |
| Grad-Cam    | 9.98 / 22.13   | 10.85 / 8.06        |
| DSS [21]    | 9.02 / 39.49   | 12.03 / 5.21        |
| EgoGaze [22]| 15.02 / 31.34  | 11.30 / 3.37        |
| Ours        | 17.43 / 48.81  | 17.86 / 29.6        |
| Ours+Obj    | 25.69 / 75.69  | 29.6 / 1.99         |

4.5. Interaction Hotspots Estimation

We now present our experimental results on interaction hotspot estimation, and compare our method against a set of baselines, including:

- **Center Prior** represents a fixed 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 [49].
- **EgoGaze** considers possible gaze position as salient region of a given image. This model is directly trained on eye fixation annotation from EGTEA-Gaze+ [22].
- **DSS Saliency** predicts salient region during human object interaction. This model is trained on pixel-level saliency annotation from [34].

The experimental results are summarized in Table 2. Among all baseline methods, EgoGaze achieves the best performance on both EGTEA and Epic-Kitchens datasets. This suggests a correlation between fixation and visual affordance, which is consistent with previous findings in the psychology literature [45]. Even so, our model further improves the F1 score by 5.4% on EGTEA and 12.6%
4.6. Motor Attention Prediction

We now report our experimental results on motor attention prediction. We consider the following baselines:

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

Predicting the hand trajectory from the first person view is a challenging task, due to the severe ego-motion and random nature of human body movement. The experimental results are presented in Table 5. Note that all baseline methods need the coordinate of the observed hand for prediction. This converts trajectory prediction into a regression problem. In contrast, our model only takes video clips as inputs and does not need any observation of hand position for inference. In addition, our model outputs a probability distribution which incorporates the uncertainty in human motion. Even so, our method only slightly lags behind the strongest LSTM baseline. However, LSTM encounters inevitable failure when the hand has not been observed, while our model is capable of “imagining” the possible hand trajectory. See “Operate Microwave” and “Wash Coffee Cup” in Fig 3. This generalization ability is attributable to our incorporation of a latent space model of motor attention. Note that motor attention here serves a vehicle towards learning the future representation. Optimization of the performance of motor attention is a topic for future research.

4.7. Analysis and Discussion

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 next-active objects and corresponding interaction hotspots. Hence, our model can address challenging cases where next-active objects are ambiguous. Take “Operate Stove” in Fig 3 as an example. The model may predict “Put Pan” without explicitly modeling motor attention. This further supports our claim that putting motor attention into the loop results in a better future representation.

One limit of our model is that modeling motor attention as probabilistic variables cannot effectively discriminate the left hand from the right hand. We leave this piece of the puzzle for future endeavours. Our model also shares a similar conundrum faced by previous anticipation studies. The model will fail when future active objects are occluded or not even observed. (See “Close Fridge Drawer” and “Put Coffee Maker” in Fig 3) This connects to a more general question in Artificial Intelligence: How can we endow an intelligent system with the ability of exploration and logical reasoning? The solution remains to be explored.
5. Conclusions

We have presented the first deep model that jointly predicts motor attention, interaction hotspots, and future action labels in FPV. We show 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 and human-object interaction. We obtain state-of-the-art action anticipation results on two FPV benchmark datasets, and strong results on motor attention and interaction hotspots estimation. We believe that our model connects findings in cognitive neuroscience to an important task in computer vision, thereby providing a solid step towards the challenging problem of visual anticipation.

References

[1] Salvatore M Aglioti, Paola Cesari, Michela Romani, and Cosimo Urgesi. Action anticipation and motor resonance in elite basketball players. Nature neuroscience, 2008. 2
[2] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In ICCV, 2019. 2, 3
[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li-Fei Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In CVPR, 2016. 6, 8
[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017. 5, 6
[5] Chao-Yeh Chen and Kristen Grauman. Subjects and their objects: Localizing interactees for a person-centric view of importance. IJCV, 2018. 3
[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 ECCV, 2018. 3, 5, 6, 7
[7] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. Scene semantics from long-term observation of people. In ECCV, 2012. 3
[8] Giuseppe Di Pellegrino, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Understanding motor events: a neurophysiological study. Experimental brain research, 1992. 1
[9] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demov2vec: Reasoning object affordances from online videos. In CVPR, 2018. 3, 8
[10] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In ICCV, 2011. 2
[11] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos. In ICCV, 2017. 2
[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In ICCV, 2015. 2, 3
[13] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. VCIP, 2017. 2
[14] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In ECCV Workshops, 2018. 7
[15] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In ICCV, 2019. 2, 5, 7
[16] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. In BMVC, 2017. 2
[17] Deepthi Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In CVPR, 2019. 7
[18] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In CVPR, 2011. 3
[19] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and Jose MF Moura. Adversarial geometry-aware human motion prediction. In ECCV, 2018. 2, 3
[20] Riitta Hari, Nina Forss, Sari Avikainen, Erika Kirveskari, Stephan Salenius, and Giacomo Rizzolatti. Activation of human primary motor cortex during action observation: a neuromagnetic study. Proceedings of the National Academy of Sciences, 1998. 1
[21] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In CVPR, 2017. 7
[22] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In ECCV, 2018. 7
[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015. 5
[24] William James, Frederick Burkhardt, Fredson Bowers, and Ignas K Skrupskelis. The principles of psychology, volume 1. Macmillan London, 1890. 1
[25] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In ICLR, 2017. 4
[26] Hirokatsu Kataoka, Yudai Miyashita, Masaki Hayashi, Kenji Itawa, and Yutaka Satoh. Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature. In BMVC, 2016. 2
[27] Qihong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In CVPR, 2019. 2
[28] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In ECCV, 2012. 2
[29] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230, 2018. 2
[30] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. TPAMI, 2015. 3
[31] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018. 2, 5, 6

[32] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *CVPR*, 2015. 2

[33] Miao Liu, Xin Chen, Yun Zhang, Yin Li, and James M Rehg. Paying more attention to motion: Attention distillation for learning video representations. *arXiv preprint arXiv:1904.03249*, 2019. 7

[34] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *TPAMI*, 2010. 7

[35] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016. 2

[36] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017. 4

[37] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. Leveraging the present to anticipate the future in videos. In *CVPR Workshops*, 2019. 2, 7

[38] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 2, 3, 6, 8

[39] Vladimir Pavlovic, James M Rehg, and John MacCormick. Learning switching linear models of human motion. In *NeurIPS*, 2001. 3

[40] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. Parsing video events with goal inference and intent prediction. In *ICCV*, 2011. 2

[41] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 2

[42] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. Compact CNN for indexing egocentric videos. In *WACV*, 2016. 2

[43] Nicholas Rhinehart and Kris M Kitani. Learning action maps of large environments via first-person vision. In *CVPR*, 2016. 3

[44] Nicholas Rhinehart and Kris M. Kitani. First-person activity forecasting with online inverse reinforcement learning. In *ICCV*, 2017. 2, 3

[45] Lucia Riggio, Cristina Iani, Elena Gherri, Fabio Benatti, Sandro Rubichi, and Roberto Nicoletti. The role of attention in the occurrence of the affordance effect. *Acta psychologica*, 127(2):449–458, 2008. 7

[46] MFS Rushworth, H Johansen-Berg, Silke Melanie Göbel, and JT Devlin. The left parietal and premotor cortices: motor attention and selection. *Neuroimage*, 20:889–900, 2003. 2

[47] MS Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *HRI*, 2015. 3

[48] Michael S Ryoo, Brandon Rothrock, and Larry Matthies. Pooled motion features for first-person videos. In *CVPR*, 2015. 2

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7

[50] Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *ECCV*, 2018. 2

[51] Suriya Singh, Chetan Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016. 6

[52] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *CVPR*, 2015. 2

[53] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don’t forget to turn the lights off! In *ICCV*, 2015. 2

[54] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. In *CVPR*, 2017. 3

[55] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 7

[56] Raquel Urtasun, David J Fleet, Andreas Geiger, Jovan Popović, Trevor J Darrell, and Neil D Lawrence. Topologically-constrained latent variable models. In *ICML*, 2008. 3

[57] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016. 2

[58] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 2, 3

[59] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *TPAMI*, 2007. 3

[60] Xiaoalong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 3

[61] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *CVPR*, 2018. 2, 3

[62] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and有害物源. Egocentric activity prediction via event modulated attention. In *ECCV*, 2018. 2

[50] Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *ECCV*, 2018. 2

[51] Suriya Singh, Chetan Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016. 6

[52] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *CVPR*, 2015. 2

[53] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don’t forget to turn the lights off! In *ICCV*, 2015. 2

[54] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. In *CVPR*, 2017. 3

[55] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 7

[56] Raquel Urtasun, David J Fleet, Andreas Geiger, Jovan Popović, Trevor J Darrell, and Neil D Lawrence. Topologically-constrained latent variable models. In *ICML*, 2008. 3

[57] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016. 2

[58] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 2, 3

[59] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *TPAMI*, 2007. 3

[60] Xiaoalong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 3

[61] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *CVPR*, 2018. 2, 3

[62] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, and Qi Tian. Cascaded interactional targeting network for egocentric video analysis. In *CVPR*, 2016. 2
Supplementary Materials

The contents of supplementary materials are organized as following:

- A Training Details and Network Architecture.
- B Mathematical Derivation for Equation 8.
- C Details on Data Annotation.
- D Full Results on the Epic-Kitchens Dataset.
- E Additional Qualitative Results.

A. Training Details and Network Architecture

In this section, we describe the training details of our model on both the EGTEA and the EPIC-Kitchens datasets. We also present the detailed network architecture. Our implementation can be found on this anonymous repository.

A.1. Implementation details

As introduced in main paper, our model takes 32 consecutive frames (subsampled by 2 in time) as the input for both EGTEA and EPIC-Kitchens datasets. This corresponds to an observation time of 2.5 seconds for EGTEA and 2 seconds for EPIC-Kitchens. We trained the model with 13D ResNet-50 backbone for 40 epochs on the EGTEA dataset and 30 epochs on the EPIC-Kitchens dataset. An interesting observation is that the EPIC-Kitchen dataset favors a shorter training time in comparison to the EGTEA dataset. A longer training schedule can slightly increase the mean class accuracy (though not evaluated by the EPIC-Kitchens Challenge), yet decrease the Top-1 accuracy. We train the CSN-152 model for even less epochs (18 epochs) on EPIC-Kitchens training set. This is because this dense model is much easier to overfit. We thus adapt an early stopping mechanism (stops at 15 epochs) to optimize the performance on the unseen kitchens.

A.2. Network Architecture

We present our network architecture in Table 6. We use the feature from shallow layer of the network for motor attention prediction and interaction hotspots estimation. The features from shallow layer can produce attention maps with higher spatial resolutions. Our model has a similar objective as previous works on multi-task learning. The key difference is that each objective of our model is highly correlated with each other. Therefore, we do not need to re-weight the total loss based on the priority of the task. In contrast, a traditional multi-task pipeline will have to increase the weights of the cross entropy loss, if the main task is action anticipation.

B. Mathematical Derivation for Equation 8

As discussed in Sec 3.5, we inject posterior \(p(A, M|x)\) into \(p(y|x)\) and optimize the resulting latent variable model by maximizing the Evidence Lower Bound (ELBO). However, the prior distribution of \(Q(A, M|x)\) is not available for training. Here, we provide additional mathematical derivation to show that minimizing \(KL[p(A, M|x)||Q(A, M|x)]\) is equivalent to minimizing \(KL[p(A|x)||Q(A|x)] + KL[p(M|x)||Q(M|x)]\).

First, we have the following conditional probability:

\[ p(A, M|x) = p(A|M, x)p(M|x). \]  

(12)

Apparently, \(p(A|M, x)\) and \(p(M|x)\) are independent. Hence, we have

\[ KL[p(A, M|x)||Q(A, M|x)] \]

\[ = KL[p(A|M, x)||Q(A|M, x)] + KL[p(M|x)||Q(M|x)] \]  

(13)

According to Eq. 5 in the main paper, the conditional probability \(p(A|x)\) can be re-written as the combination of \(p(A|M, x)\) and \(p(M|x)\). Therefore, minimizing \(KL[p(A, M|x)||Q(A, M|x)]\) is indeed equivalent to minimizing \(KL[p(A|x)||Q(A|x)] + KL[p(M|x)||Q(M|x)]\).

C. Details on Data Annotation

In Sec. 4.1, we introduced how we obtain the prior distribution of motor attention. Here we show a visual illustration of the approximation process of future hand position on the EPIC-Kitchens dataset in Fig. 4 (a). We also present more details about the interaction hotspots annotation process. An example can be found in Fig. 4 (b).
For each sample, we compare the last observable frame with the first frame of action segment. If the active object presents in the last observable frame, we annotate the corresponding contact point and enforce a 2D Gaussian distribution to imitate the uncertainty of human-object interaction. If the active objects is missing from the last observable frame, we assume a uniform distribution during training. To summarize, there are 14951 annotated sample on the EPIC-Kitchens Dataset, and 7381 annotated samples on the EGTEA dataset. Note that we use a smaller anticipation time (0.5s) on the EGTEA dataset. This is because the EGTEA dataset has a smaller angle of view in comparison with the EPIC-Kitchens dataset. A large anticipation time will reduce the number of samples that have next-active objects on the last observable frame.

D. Full Results on the Epic-Kitchens Dataset.

Fig 5 presents a screenshot of the leaderboard from the EPIC-Kitchens Egocentric Action Anticipation Challenge (https://epic-kitchens.github.io/). The screenshot was acquired on the end date of Phase 2 challenge (2019.11.22). To date, our proposed method outperforms all published results by a large margin. Several unpublished works (user id: “action_banks”, “reza_zlf”, “hepic”, “prefact” in Fig 5) also attempt at the EPIC-Kitchen Anticipation Challenge. On the seen kitchens (S1), “action_banks” slightly outperforms our method for action prediction, but it is inferior to our method in terms of the verb and noun prediction. On the unseen kitchens (S2), our method outperforms “action_banks” for all anticipation tasks by a notable margin.

E. Additional Qualitative Results

Finally, we provide additional qualitative results. The video demo included in the supplementary materials demonstrates our results. Here we illustrate more samples of predicted motor attention, interaction hotspots, and action labels in Fig 6. The figure follows the same format as Fig. 3 in the submission. These results further show that our proposed motor attention module has the remarkable ability of “imagining” possible hand movements even without the presence of hands in the observed video segments. Another interesting observation is that the predicted distribution of interaction hostpots can be sparse in certain circumstances (e.g., “Open Fridge” or “Take Condiment”). This is because human-object interaction is a stochastic process. There might be multiple valid contact regions for manipulation, especially when the next-active object has a relatively large scale. This again shows the necessity of the stochastic units in our proposed method.

As discussed in our main paper, the occlusion and absence of active objects make the anticipation problem intractable even for humans. The failure cases in Fig. 3 also suggest that the anticipation model can be biased by ongoing action. This is because current FPV datasets (especially EPIC-Kitchens) segment a continuous action into several same subatom actions to ensure all action segments have similar temporal dimension. For instance, A video clip of “cutting onions” for 20 seconds is segmented into 7 or 8 shorter clips all having the same “cutting onions” label. This increases the transition probability of staying in current state, and thereby biases the model. Therefore, the ability of predicting when exactly the action will end is important for more accurate action prediction model. This task is also related to the action localization problem in the literature.
Figure 5. Screenshot from Epic-Kitchens Anticipation Challenge. The user name of our proposed method is “aptx4869lm”. Note that user “antonionfurnari” refers to RULSTM in our main paper. They further improved the results reported in their paper.
Figure 6. Additional visualization of predicted motor attention (left), interaction hotspots (right), and future action labels (top) from the EGTEA dataset (1-4 row) and the EPIC-Kitchens dataset (5-8 row). Both successful cases (green label) and failure cases (red label) are presented. Future hands position are downsampled by a temporal factor of 8, and forecasted to the last observable frame in the order of yellow, green, cyan, and magenta.
| ID | Branch | Type | Kernel Size THWC (C) | Stride THW | Output Size THWC | Comments (Loss) |
|----|--------|------|----------------------|-----------|-----------------|-----------------|
| 1  |        | Conv3D | 5x7x7,64             | 2x2x2     | 16x112x112x64  |                 |
| 2  |        | MaxPool1 | 2x3x3             | 2x2x2     | 8x56x56x64     |                 |
| 3  | Layer1 Bottleneck 0-2 | 3x1x1,64 | 1x3x3,64 (3 times) 1x1x1,256 | 1x1x1 (3 times) 1x1x1 | 8x56x56x256 | Addition Pooling |
| 4  |        | MaxPool2 | 2x1x1             | 2x1x1     | 4x56x56x256   | Reduce Memory Usage |
| 5  | Layer2 Bottleneck 0 | 3x1x1,128 | 1x3x3,128 1x1x1,512 | 1x1x1 1x2x2 (3 times) 1x1x1 | 4x28x28x512 |
| 6  | Backbone (shared) | Layer2 Bottleneck 1-3 | 3x1x1,128 | 1x3x3,128 (3 times) 1x1x1,512 | 1x1x1 1x2x2 (3 times) 1x1x1 | 4x28x28x512 |
| 7  |        | Layer3 Bottleneck 0 | 3x1x1,256 | 1x3x3,256 1x1x1,1024 | 1x1x1 1x2x2 1x1x1 |                 |
| 8  | Layer3 Bottleneck 1-5 | 3x1x1,256 | 1x3x3,256 (5 times) 1x1x1,1024 | 1x1x1 1x1x1 (5 times) 1x1x1 | 4x14x14x1024 |
| 9  | Layer4 Bottleneck 0 | 3x1x1,128 | 1x3x3,128 1x1x1,512 | 1x1x1 1x2x2 1x1x1 |                 |
| 10 | Layer4 Bottleneck 1-2 | 3x1x1,128 | 1x3x3,128 (2 times) 1x1x1,512 | 1x1x1 1x2x2 (2 times) 1x1x1 | 4x7x7x2048 |
| 11 | Motor Attention Module | Conv3d 1 (on Layer 2 feature) | 1x3x3,128 | 1x1x1     | 4x28x28x128  |                 |
| 12 |        | Conv3d 2 | 1x3x3,1             | 1x1x1     | 4x28x28x1     | KLD Loss        |
| 13 |        | Maxpool 1 | 1x2x2             | 1x2x2     | 4x14x14x1     | Guiding Interaction Hotspots |
| 14 |        | Gumbel Softmax 1 (Sampling) | 1x3x3,128 | 1x1x1 | 4x14x14x1 | Sampling Motor Attention |
| 15 |        | Maxpool 2 | 1x1x1,1024       | 1x1x1     | 4x14x14x1     | Guiding Action Anticipation |
| 16 |        | Gumbel Softmax 2 (Sampling) | 1x3x3,128 | 1x1x1 | 4x14x14x1 | Sampling Motor Attention |
| 17 | Interaction Hotspots Module | Conv3d 1 (on Layer 3 Feature) | 1x3x3,256 | 1x1x1 | 4x14x14x256 | With Sampled Motor Attention |
| 18 |        | Conv3d 2 | 1x3x3,1             | 1x1x1     | 4x14x14x1     | KLD Loss        |
| 19 |        | Maxpool 1 | 1x2x2             | 1x2x2     | 4x14x14x1     | Guiding Action Anticipation |
| 20 |        | Gumbel Softmax (Sampling) | 1x3x3,128 | 1x1x1 | 4x14x14x1 | Sampling Interaction Hotspots |
| 21 | Action Anticipation Module | Weighted Avg Pool (on Final Feature) | 4x7x7 | 4x7x7 | 1x1x1x1024 | With Sampled Motor Attention and Interaction Hotspots |
| 22 |        | Fully Connected | 1x1x1x1N            | 1x1x1N     | 4x7x7x1     | Cross Entropy Loss (Action Anticipation) |
| 23 |        | Softmax | 4x7x7             | 4x7x7     | 1x1x1x1024 |                 |

Table 6. Network architecture of our proposed model. We omit the residual connection in backbone ResNet-50 for simplification.