Multimodal Interactive Fusion for Action Anticipation

Author Names Omitted for Anonymous Review. Paper-ID [6]

***Abstract*—Action anticipation is targeted to estimate the correct *verb-noun* that describes a human's intended movements (*verb*) and objects (*noun*) in the future from the hundreds of possible *verb-noun* combinations. Different from existing methods using commonly-used fusion mechanisms (e.g., early fusion, late fusion and hybrid fusion) or transformer-based fusion mechanism (i.e., assigning a certain modal feature as *query* and other modal features as *key* and *value*) to fuse multimodal features, this paper for the first time proposes the *Multi-round Interactive Fusion* (*MIF*) mechanism for action anticipation. Specifically, we propose two kinds of *MIF* mechanisms, namely *Parallel MIF* and *Progressive MIF*. *Parallel MIF* treats each modal feature equally and each modal feature is alternately to be assigned as the *query*. *Progressive MIF* firstly fuses the features conveying the static information (i.e., scene and object features), and then combines the dynamic information (i.e., optical flow) to generate the final feature. Benefitting from the proposed *MIF*, our model outperforms state-of-the-art methods by large margins on two public datasets, achieving 14.07% action accuracy improvement on the EGTEA dataset compared with recently proposed AFFT model.**

I. Introduction

With the rapid development of artificial intelligence technologies, anticipating human actions (as illustrated in Figure 1) is essential in many areas. However, human actions exhibit the great diversity and can be influenced by various factors. This complexity demands prediction models capable of dealing with uncertainty and possessing high generalization abilities. Moreover, human actions exhibit complex time dependence, requiring models to consider more contextual information.

To accurately anticipate actions, researchers have proposed a number of methods [3, 4, 7, 8, 10, 12, 14]. They typically fuse three modes of data: RGB, optical flow (FLOW) and object (OBJ). The fusion of multimodal data improves the prediction ability for several reasons: Multimodal fusion provides a more comprehensive perspective by capturing relationships among scenes, objects and motions, which helps to better understand the contextual information of actions. Although these methods have achieved some success, there are still limitations. Currently, the main fusion methods used are *early fusion*, *late fusion* and *hybird fusion*, they are relatively simple and rough. *Early fusion* can learn to exploit the correlation and interactions between low-level features of each modality. *Late fusion* refers to a method in which a model is trained on different modalities and then the outputs are fused [1]. *Hybrid fusion* combine *early fusion* and *late fusion*, integrating the advantages of *early fusion* to capture feature relationships and *late fusion* to handle overfitting [16], but also increases the training difficulty of the model. As a result, models produced by these fusion methods have some obvious shortcomings, including: 1) inadequate processing of multimodal data leading to characteristic information loss and repetition, 2) insufficient correlation modeling to capture complex relationships across modalities and 3) lack of flexibility and adaptability affects the robustness and reliability of models.

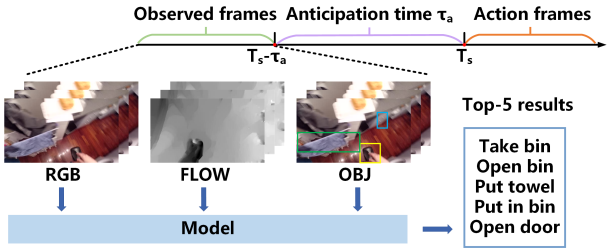

Fig. 1. The action anticipation task predicts the next action based on the observed frames. Anticipation time τ_a is how much in advance the action has to be anticipated.

Therefore, this paper proposes a ***M**ulti-round **I**nteractive **F**usion* (***MIF***) model to address the aforementioned shortcomings. ***MIF*** utilizes interactions between different modalities to facilitate information exchange and integration, providing a better way to capture semantic information and relationships across modalities. Furthermore, it improves the efficiency of data training and information mining. We explore two different interactive fusion methods. One is the parallel interactive fusion, and the other is progressive interactive fusion. Each has its own advantages. To verify the validity of our model, we compare our ***P**arallel **MIF*** and ***P**rogressive **MIF*** with existing methods on EPIC-Kitchens [2] and EGTEA Gaze+ [9]. The experimental results show that the two proposed fusion methods achieve significant performance improvement in action anticipation.

The main contributions of this paper are as follows: 1) we propose two kinds of ***MIF*** mechanisms, both of them can effectively integrate information from different modes and achieve accurate action prediction, 2) we explore the combination of different modal features and verify the influence of various combinations on the model performance, this could help other researchers make more effective decisions and design models when dealing with similar problems and 3) our model has simple structure and high precision.

II. Method

A. Stage 1 Single modality feature strengthen

The modal features contain different information. Directly fusing information from these modalities may cause confusion or information loss, so it is necessary to first strengthen single-modality features before fusion. Therefore, we independently

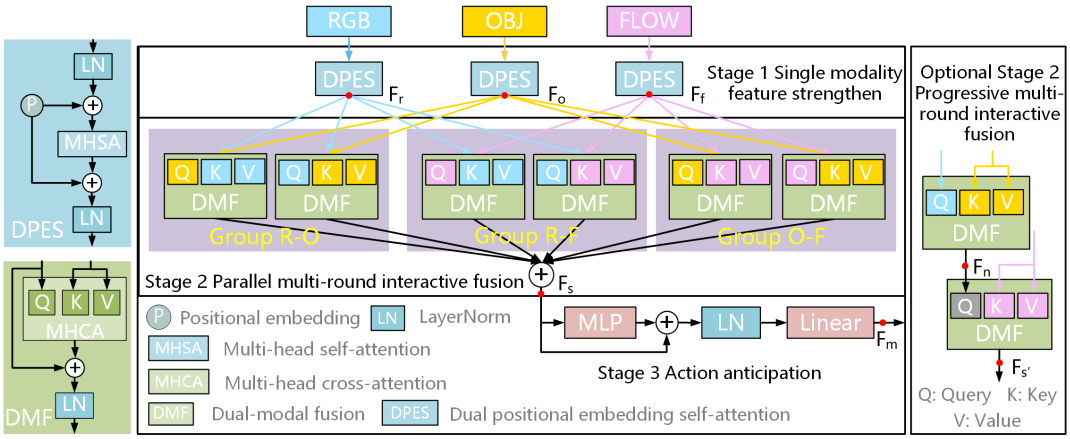

Fig. 2. The structure of ***Multi-round Interactive Fusion*** (***MIF***).

calculate attention for each modality to maintain distinct information between each modality. For this aim, we propose the dual positional embedding self-attention ($\mathcal{DPES}$) module, which embeds the positions twice and it can maintain distinct information to enhance the model's timing processing ability. Stage 1 consists of three $\mathcal{DPES}$, and it can be represented by the following formula:

$$\boldsymbol{F_i} = \mathcal{DPES}(\boldsymbol{X_i}) \ \ where \ i = r, o, f \tag{1}$$

The inputs $\boldsymbol{X_i}$ represent the initial features of modality i. The outputs $\boldsymbol{F_i}$ are the features processed by $\mathcal{DPES}$. r, o, f stand for RGB, OBJ and FLOW, respectively. In the DPES module, the computation of $\mathcal{MHSA}$ follows its original authors [13].

B. Stage 2 *Parallel* ***MIF***

The existing modal fusion methods ignore the differences and correlations between different modalities, this may result in missing information and redundant information in the model. Therefore, we propose two ***MIF*** mechanisms. In this section, we first introduce ***Parallel MIF*** (as illustrated in Figure 2), which assign equal status to the three modalities, enabling them to take turns providing *query* and attending to the information from each modality. The bidirectional correlation structures of ***Parallel MIF*** can make the model better learn the relationships between two modalities and has the ability of error correction because each group in Stage 2 exhibits bidirectional correlation. If an error occurs in one direction, the other direction can be corrected. For instance, in the two $\mathcal{DMF}$ modules of *Group R-O*, the roles of $\boldsymbol{F_o}$, and $\boldsymbol{F_r}$ are reversed.

$\mathcal{DMF}$ is based on multi-head cross-attention ($\mathcal{MHCA}$). We fuse the two modalities in such a way that one modal features provide *query* to $\mathcal{MHCA}$ while the other modal features provide *key* and *value*. The inputs and outputs of $\mathcal{MHCA}$ are given by the following formula:

$$\boldsymbol{F_{ij}} = \mathcal{MHCA}(\boldsymbol{F_i}, \boldsymbol{F_j}) \ \ where \ i, j \in \{r, o, f\} \ and \ i \neq j \tag{2}$$

where $\boldsymbol{F_{ij}}$ are the fusion features of modalities i and j. $\boldsymbol{F_i}$ provide Q, $\boldsymbol{F_j}$ provide K and V. The calculation of $\mathcal{MHCA}$ is similar to $\mathcal{MHSA}$, except that the providers of Q, K, and V features are not the same.

After obtained the outputs $\boldsymbol{F_{ij}}$ of $\mathcal{MHCA}$, we add residual and LayerNorm to it to get the outputs $\boldsymbol{F_{ij}^{'}}$ of $\mathcal{DMF}$.

$$\mathcal{DMF}(\boldsymbol{F_i}, \boldsymbol{F_j}) = LN(\boldsymbol{F_{ij}} + \boldsymbol{F_i}) = \boldsymbol{F_{ij}^{'}} \tag{3}$$

Finally we sum the outputs of the six $\mathcal{DMF}$ modules to get the final value:

$$\boldsymbol{F_s} = \sum \boldsymbol{F_{ij}^{'}} \tag{4}$$

The ***Parallel MIF*** has a bidirectional error correction ability, resulting in strong robustness. It can adapt well to different datasets and does not require us to consider which modal data has a strong modal correlation.

C. Optional Stage 2 *Progressive* ***MIF***

We provide an Optional Stage 2 ***Progressive MIF***, which can replace Stage 2. ***Parallel MIF*** and ***Progressive MIF*** have their own advantages. ***Parallel MIF*** is more robust and less thoughtful. ***Progressive MIF*** is more flexible and can achieve higher precision.

In this stage, we take into account the differences among each modality and assign different weights to the information they provide. Firstly, we fuse the static features $\boldsymbol{F_r}$ and $\boldsymbol{F_o}$ by utilizing $\boldsymbol{F_r}$ as the primary source of information for the *query* and residual. During this process, $\boldsymbol{F_r}$ play the most crucial guiding roles. Then, we deploy the fused static features $\boldsymbol{F_n}$ as the *query* to combine dynamic features $\boldsymbol{F_f}$. The process is as follow formula:

$$\mathcal{DMF}(\boldsymbol{F_r}, \boldsymbol{F_o}) = \boldsymbol{F_n} \tag{5}$$

$$\mathcal{DMF}(\boldsymbol{F_n}, \boldsymbol{F_f}) = \boldsymbol{F_{s'}} \tag{6}$$

Given that $\boldsymbol{F_r}$ and $\boldsymbol{F_o}$ are static features and possess high correlation due to both representing object-related information. After obtaining the static fused features $\boldsymbol{F_n}$ that combine rich information, we use $\boldsymbol{F_n}$ to combine dynamic features $\boldsymbol{F_f}$ and fuse them to complement each other's information.

D. Stage 3 Action anticipation

In the Stage 3 Action anticipation, the outputs are final anticipation results. The following formula serves as an example of using $\boldsymbol{F_s}$ as inputs for Stage 3.

$$\boldsymbol{F_m} = Linear(LN(MLP(\boldsymbol{F_s}) + \boldsymbol{F_s})) \tag{7}$$

The multi-layer perceptron increases the number of trainable weights, and its hierarchical representation learning improves

TABLE I
THE COMPARISON RESULTS WITH THE STATE-OF-THE-ART.

Method	EPIC-Kitchens							EGTEA Gaze+						
	Top-5 Acc			M.Top-5 Rec			Avg.	Top-5 Acc			M.Top-5 Rec			Avg.
	VERB	NOUN	ACT	VERB	NOUN	ACT		VERB	NOUN	ACT	VERB	NOUN	ACT	
ATSN[2] ECCV'2018	77.30	39.93	16.29	33.08	32.77	7.60	34.50	90.60	69.94	40.53	69.24	57.02	31.61	59.82
FN[6] WACV'2018	74.84	40.87	26.27	35.30	37.77	6.64	36.95	91.05	71.64	60.12	76.73	63.59	49.82	68.83
RULSTM[4] ICCV'2019	79.55	51.79	35.32	43.72	49.90	15.10	45.90	93.11	77.48	66.40	82.07	73.30	58.64	75.17
TAB[12] ECCV'2020	79.47	51.93	34.60	44.15	51.88	16.17	46.37	92.84	78.58	67.51	83.12	75.17	62.46	76.61
IRNN[14] TIP'2020	79.70	50.20	33.20	-	-	-	-	-	-	-	-	-	-	-
HA[8] 2021	73.94	41.29	25.18	35.87	36.39	14.64	37.89	86.06	71.46	59.89	78.90	69.89	58.22	70.74
HRO[10] CVPR'2022	81.53	54.51	37.42	45.16	51.78	17.80	48.03	-	-	71.46	-	-	-	-
Parallel MIF	80.89	57.38	39.40	47.37	57.65	17.90	50.10	96.23	88.28	79.89	90.74	85.98	75.70	86.14
Progressive MIF	80.93	57.76	39.80	47.99	59.04	17.80	50.55	96.24	88.24	80.22	90.74	86.24	76.54	86.37

TABLE II
THE COMPARISON RESULTS WITH DIFFERENT FUSION STRATEGIES.

Method	EPIC-Kitchens							EGTEA Gaze+						
	Top-5 Acc			M.Top-5 Rec			Avg.	Top-5 Acc			M.Top-5 Rec			Avg.
	VERB	NOUN	ACT	VERB	NOUN	ACT		VERB	NOUN	ACT	VERB	NOUN	ACT	
Early fusion	80.27	50.46	33.27	45.47	51.38	14.35	45.87	96.11	87.95	79.61	90.19	85.42	75.43	85.79
Late fusion	80.09	55.31	39.06	41.01	53.50	14.21	47.20	95.35	85.64	78.57	86.62	81.08	70.62	82.98
Hybrid fusion (r-o)	80.93	52.03	36.63	44.34	50.01	14.16	46.35	-	-	-	-	-	-	-
Hybrid fusion (r-f)	80.73	56.60	38.94	43.25	56.49	15.53	48.59	-	-	-	-	-	-	-
Hybrid fusion (f-o)	80.07	48.81	33.19	40.71	46.60	11.22	43.43	-	-	-	-	-	-	-
Parallel MIF	80.89	57.38	39.40	47.37	57.65	17.90	50.10	96.23	88.28	79.89	90.74	85.98	75.70	86.14
Progressive MIF	80.93	57.76	39.80	47.99	59.04	17.80	50.55	96.24	88.24	80.22	90.74	86.24	76.54	86.37

the model's generalization ability. Furthermore, the residual structure preserves the original information and accelerates convergence during the training process.

III. EXPERIMENTS

A. Experiment setup

Datasets. EPIC-Kitchens (EPIC) [2] is a dataset that contains 125 verb classes and 352 noun classes. EGTEA Gaze+ (EGTEA) [9] has 19 verb classes, 51 noun classes and 106 action classes. EGTEA is official split into three parts. We report the average of the three splits. For fairness, we use the pre-extracted modal features by [4] in the two datasets and conduct tests on validation sets which are split by [4].

Fusion details. EGTEA only provides two modalities: RGB and FLOW, so there are only two DMF modules in Stage 2 on EGTEA, and in the two modules, $\boldsymbol{F_r}$ and $\boldsymbol{F_f}$ take turns to provide *query*. In Optional Stage 2, we use $\boldsymbol{F_r}$ provide the *query* of the first DMF module and the *key* and *value* of the second DMF module.

B. Comparisons to previous works

Comparisons with the state-of-the-art. We compare our model to the state-of-the-art in Table I, and the results demonstrate that our model achieves competitive performance. The anticipation time τ_a is 1s. 'Avg.' represents the average value of all metric results for each method. Most existing methods employ late fusion and do not thoroughly investigate the relationships between features from different modes. Features across these modes contain complementary information, and we integrate them interactively to obtain richer features representation and a more comprehensive understanding of the scene. Among our two methods, '***Progressive MIF***' shows superior overall performance compared to '***Parallel MIF***'. However, '***Parallel MIF***' still has its advantages and significance. The parallel method is more robust and assigns equal importance to all modalities, eliminating the need to consider which modality should be given more weight, even when applied to other datasets.

Comparisons with different fusion strategies. To verify that interactive fusion is more effective than other fusion methods, we conducted comparisons with different fusion strategies and the results are shown in Table II. For fairness, during the comparisons, other structures of our model remained unchanged. In the table, '*Hybrid fusion (r-o)*' denotes the two features in parentheses undergoing *early fusion* and the remaining feature being fused later. As shown in the table, our methods achieve the best results. Compared to the other methods, ***MIF*** is simpler and more effective. In our methods, features are fused at a higher level to capture richer semantic information compared with *early fusion*. Our methods better capture complex relationships between different modal features compared with *late fusion*, leading to improved generalization ability. And our model is not like *hybrid fusion*, where the choice of strategy and parameters has a great impact on the result.

Comparisons with different anticipation time τ_a. The results in Table III demonstrate that our model significantly

TABLE III
THE TOP-5 ACTION ACCURACY (%) RESULTS ON DIFFERENT ANTICIPATE TIME τ_a.

Method	EPIC-Kitchens					EGTEA Gaze+				
	2	1.75	1.5	1.25	1	2	1.75	1.5	1.25	1
ED[5] BMVC'2017	21.53	22.22	23.20	24.78	25.75	45.03	46.22	46.86	48.36	50.22
FN[6] WACV'2018	23.47	24.07	24.68	25.66	26.27	54.06	54.94	56.75	58.34	60.12
RU-LSTM[4] ICCV'2019	29.44	30.71	32.33	33.41	35.32	56.82	59.13	61.42	63.53	66.40
SRL[11] TPAMI'2021	30.15	31.28	32.36	34.05	35.52	59.69	61.79	64.93	66.45	70.67
HRO[10] CVPR'2022	31.30	32.67	34.26	35.87	37.42	60.12	62.32	65.53	67.18	71.46
Parallel MIF	39.24	39.32	39.02	38.82	39.40	79.73	79.70	79.51	79.94	79.94
Progressive MIF	39.58	39.58	39.28	39.26	39.80	79.90	79.91	79.87	80.15	80.22

outperforms existing methods for various anticipation times. Unlike previous methods, our method's action accuracy does not decrease as the anticipation time lengthens. This suggests that our model is better suited for long-term anticipation. The main reason is the complementarity of information from various modalities in interactive fusion. Interactive fusion is interactively fused at different information level, and it can make full use of multimodal information to reduce the error rate of model attention. So that they are not susceptible to the expected time, even for long-term predictions, the model's accuracy does not decline substantially.

TABLE IV
THE TOP-1 ACTION ACCURACY (%) RESULTS ON EGTEA, τ_a IS 0.5S.

Method	Top-1 Acc		
	VERB	NOUN	ACT
AVT[7] ICCV'2021	51.70	50.30	39.80
AFFT[15] WACV'2023	53.40	50.40	42.50
Parallel MIF	58.36	59.00	47.99
Progressive MIF	58.16	59.25	48.48

Comparisons of the top-1 action accuracy. In Table IV, we present the top-1 accuracy results at $\tau_a = 0.5$ on the EGTEA dataset, indicating that our model still has a significant advantage in top-1 action accuracy over previous methods.

C. Ablation Studies

In order to verify the rationality and effectiveness of the parallel interactive fusion structure, we conducted ablation experiments on it. Table V shows the action performance results of different multimodal parallel interactive fusion manners. '$\boldsymbol{F_r}$-Q', '$\boldsymbol{F_f}$-Q' and '$\boldsymbol{F_o}$-Q' are refer to the $\mathcal{DMF}$ modules in Stage 2 with only $\boldsymbol{F_r}$, $\boldsymbol{F_f}$ and $\boldsymbol{F_o}$ provide *query*, respectively. 'Diff-All' means Q, K and V are different values and selected from $\{\boldsymbol{F_r}, \boldsymbol{F_o}, \boldsymbol{F_f}\}$. One of the reasons ***Parallel MIF*** performs better is due to the bidirectional correlation between modes. In contrast, for '$\boldsymbol{F_i}$-Q' ($i = r, o, f$), the correlation is unidirectional, which may result in important information being ignored. The ability to correct errors may be limited. For EPIC, we have an additional reason because EPIC has three modes data. The reason is as follow: when $\boldsymbol{F_r}$ provide *query*, model does not explore the relationships between $\boldsymbol{F_o}$ and $\boldsymbol{F_f}$, leading to weakened the interactive fusion of static and dynamic features. Regarding 'Diff-All', we analyze that the result of 'Diff-All' is not as good as that of our method because our method handles the correlation between the two modes independently. In 'Diff-All', *query*, *key* and *value* are provided from different modes, which may confuse the model when learning the relationships between modes.

TABLE V
THE ABLATIONS OF ***Parallel MIF***.

Manner	EPIC-Kitchens		EGTEA Gaze+	
	Top-5 Acc	M.Top-5 Rec	Top-5 Acc	M.Top-5 Rec
F_r-Q	38.13	15.60	79.74	75.56
F_f-Q	38.94	16.32	80.08	75.45
F_o-Q	38.42	16.73	-	-
Diff-All	38.70	16.45	-	-
Parallel MIF	39.40	17.90	80.22	76.54

TABLE VI
THE ABLATIONS OF ***Progressive MIF*** ON EPIC.

Manner	Top-5 Acc	M.Top-5 Rec
$\langle O, R\rangle \rightarrow F$	38.07	17.27
$\langle O, F\rangle \rightarrow R$	38.90	16.48
$\langle F, R\rangle \rightarrow O$	39.12	16.77
$\langle F, O\rangle \rightarrow R$	38.52	17.55
$\langle R, F\rangle \rightarrow O$	39.40	17.06
$\langle R, O\rangle \rightarrow F$	39.80	17.80

In order to verify the rationality and effectiveness of ***Progressive MIF*** , we conducted ablation experiments on it. Table VI presents the results of the action performance using different multimodal progressive interactive fusion manners on the EPIC. $\langle X, Y\rangle \rightarrow Z$ means in the first $\mathcal{DMF}$ module, the features of X provide *query*, the features of Y provide *key* and *value*. In the second $\mathcal{DMF}$ module, *key* and *value* obtained by Z. In Table VI, the results of $\langle R, O\rangle \rightarrow F$ are better than others. We think the advantage of this design is that $\boldsymbol{F_o}$ has stronger correlation with $\boldsymbol{F_r}$, so learning the relationships between $\boldsymbol{F_r}$ and $\boldsymbol{F_o}$ first helps to capture more effective static information. Moreover, as $\boldsymbol{F_r}$ contains most of the relevant information, $\boldsymbol{F_r}$ are used as the *query* in the first $\mathcal{DMF}$ module, achieving the highest scores.

IV. CONCLUSION

In this paper, we propose two simple structured ***MIF*** mechanisms, ***Parallel MIF*** and ***Progressive MIF***. The ***Parallel MIF*** possesses bidirectional correlation structures, which endows it with stronger robustness. In contrast, the ***Progressive MIF*** use static features to combine dynamic features to achieve higher accuracy. We compare our proposed methods to the state-of-the-art and other multimodal fusion methods, and the experimental results show that our proposed methods achieve superior performance on two large-scale datasets.

ACKNOWLEDGMENTS

REFERENCES

[1] G. Castellano, L. Kessous, and G. Caridakis. Emotion recognition through multiple modalities: Face, body gesture, speech. *Springer Berlin Heidelberg*, 2008.

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, editor=Ferrari Vittorio Wray, Michael, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Scaling egocentric vision: The dataset. In *European Conference on Computer Vision*, 2018.

[3] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13219–13228, 2021.

[4] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision*, 2019.

[5] J. Gao, Z. Yang, and R. Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. In *British Machine Vision Conference*, 2017.

[6] R De Geest and T. Tuytelaars. Modeling temporal structure with lstm for online action detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1549–1557, 2018.

[7] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. *IEEE/CVF International Conference on Computer Vision*, pages 13485–13495, 2021.

[8] X. Gu, J. Qiu, Y. Guo, B. Lo, and G. Z. Yang. Transaction: Icl-sjtu submission to epic-kitchens action anticipation challenge. 2021.

[9] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision*, 2018.

[10] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13903, 2022.

[11] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv*, 2017.

[14] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020.

[15] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6057–6066, 2023.

[16] Zhen zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G. Hauptmann. Multimedia classification and event detection using double fusion. *Multimedia Tools and Applications*, 71(1):333–347, 1 2014.