Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities

Michele Mazzamuto[1*], Antonino Furnari[1*], Yoichi Sato[2], Giovanni Maria Farinella[1]
[1]University of Catania, Catania, Italy
michele.mazzamuto@phd.unict.it, antonino.furnari@unict.it, giovanni.farinella@unict.it
[2]The University of Tokyo, Tokyo, Japan
ysato@iis.u-tokyo.ac.jp
*Co-first authors.

Abstract

We address the challenge of unsupervised mistake detection in egocentric video of skilled human activities through the analysis of gaze signals. While traditional methods rely on manually labeled mistakes, our approach does not require mistake annotations, hence overcoming the need of domain-specific labeled data. Based on the observation that eye movements closely follow object manipulation activities, we assess to what extent eye-gaze signals can support mistake detection, proposing to identify deviations in attention patterns measured through a gaze tracker with respect to those estimated by a gaze prediction model. Since predicting gaze in video is characterized by high uncertainty, we propose a novel gaze completion task, where eye fixations are predicted from visual observations and partial gaze trajectories, and contribute a novel gaze completion approach which explicitly models correlations between gaze information and local visual tokens. Inconsistencies between predicted and observed gaze trajectories act as an indicator to identify mistakes. Experiments highlight the effectiveness of the proposed approach in different settings, with relative gains up to $+14\%$, $+11\%$, and $+5\%$ in EPIC-Tent, HoloAssist and IndustReal respectively, remarkably matching results of supervised approaches without seeing any labels. We further show that gaze-based analysis is particularly useful in the presence of skilled actions, low action execution confidence, and actions requiring hand-eye coordination and object manipulation skills. Our method is ranked first on the HoloAssist Mistake Detection challenge.

1. Introduction

Smart glasses are gaining more and more popularity, with various existing products capable of supporting the user

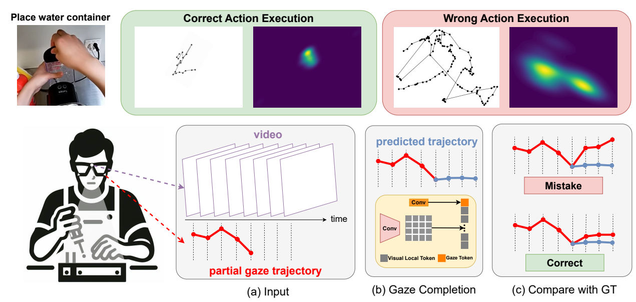


Figure 1. Top: gaze trajectories of a correct and wrong execution of the "place water container" action, together with gaze fixation maps averaged across many action instances. Note the higher variability exhibited by wrong executions. Bottom: (a) The proposed unsupervised mistake detection method assumes as input a video with a partial gaze trajectory on the initial part of the video. (b) A gaze completion model predicts a gaze trajectory for the remaining part of the video, conditioned on the input video and the partial trajectory. (c) A mistake is detected if the predicted trajectory is significantly different from the observed one, suggesting a deviation from the expected attention patterns.

through Augmented Reality. In order to provide timely assistance, wearable devices should be able to identify moments in which the user makes mistakes or is confused and requires help [7]. If such instances are properly detected, the AI system can proactively offer contextual information or suggestions on how to best carry out the task at hand [42].

Previous works tackled the problem of detecting mistakes from a fully supervised perspective, where mistake instances were labeled in egocentric video and machine learning algorithms were trained to discriminate between video segments of correct action executions and incorrect ones [34, 42]. Such a fully supervised approach has two main downsides: 1) It is domain-dependent, hence requiring an accurate characterization of what a mistake is, depending on the context (e.g., a mistake in a kitchen is different from a mistake in the assembly line); 2) It requires to collect and label a sufficient number of mistake instances, which may be difficult to observe and record, involve time consuming procedures, and require expert knowledge. Another class of methods [9, 33] aim to detect mistakes without relying on mistake annotations, but still requires domain-specific and costly temporal action annotations. Ideally, a wearable assistant should be able to infer when the behavioral patterns of the user deviate from the norm in order to determine if they need assistance in a scenario-independent setting, i.e., without making specific assumptions on how a mistake is defined and without requiring costly labels. To overcome the aforementioned limitations, we propose to detect mistakes in egocentric videos of human activity in an unsupervised way, learning from unlabeled video.

Mistakes in task execution, in particular for tasks requiring hand-eye coordination and object manipulation skills, often involve abnormal attention patterns of the camera wearer [17, 18]. For instance, imagine a user operating a coffee machine without first adding water. As they press the *brew* button, they notice that no coffee is produced and start shifting their attention erratically between the cup, the water tank, and the LED indicator, deviating from the typical attention sequence of "button $\rightarrow$ cup $\rightarrow$ button." (See Figure 1(top)). This behavior is well-documented in psychology literature, which shows that gaze patterns are crucial for the execution of even the most repetitive daily activities (e.g., making tea) [20], and that they change in response to task complexity [30] and mistakes [29].

Following these observations, we study how the analysis of eye-gaze fixations can support mistake detection in egocentric video of skilled human activities. We hence propose to learn a model of "normal" attention patterns in the form of a gaze predictor producing likely eye gaze trajectories from a video at inference time. Since gaze prediction can be governed by high uncertainty, depending on the user's goals, we propose a novel "gaze completion" task in which a model takes as input a video and a partial gaze trajectory (Figure 1(a)) and is tasked to predict a likely continuation of the partial trajectory (Figure 1(b)). Gaze completion is tackled with a novel approach based on a Gaze-Frame Correlation module which explicitly models the correlation between gaze information and each local visual token. We expect videos of correct action executions to represent normal user behavior, and hence to be characterized by predictable gaze patterns, while human behavior will deviate from normality, and gaze will be unpredictable, when a mistake is made by the user. We hence signal a mistake by comparing the predicted gaze with the ground truth eye gaze trajectory obtained through a gaze tracker (Figure 1(c)).

Experiments show the effectiveness of the proposed approach, both alone, or in combination with other techniques, when compared to one-class anomaly detection methods [8, 39], and various unsupervised mistake detection baselines, with relative gains up to $+14\%$, $+11\%$, and $+5\%$ in the on EPIC-Tent [15], HoloAssist [42] and IndustReal [32] datasets respectively, remarkably matching the results of supervised methods without any labels in one-class settings. Our analysis also shows that gaze is most effective in the presence of complex actions, low-confidence executions, and actions requiring hand-eye coordination and object-manipulation skills.

In summary, the contributions of this work are as follows: 1) We investigate for the first time the problem of unsupervised mistake detection from egocentric video of human activity and provide an initial benchmark based on three datasets. 2) We define the novel "gaze completion" task where models predict gaze trajectories from video and partial gaze inputs, and introduce an approach based on a Gaze-Frame Correlation module; 3) We propose an approach to unsupervised mistake detection leveraging gaze completion to identify instances of unpredicable gaze patterns. Experiments analyze under which conditions gaze-based analysis is most useful and show the effectiveness of the approach, in one-class and unsupervised settings.

We will publicly release the code and model checkpoints to support future research.

2. Related Work

Egocentric gaze estimation Literature on gaze estimation from egocentric video is rich, with previous works investigating simultaneous gaze prediction and action recognition [5], describing gaze prediction approaches incorporating egocentric cues [21], modeling task-dependent attention transition [13], leveraging vanishing point, manipulation point, hand regions [41], introducing specific architectures [1, 19], and proposing datasets to study egocentric gaze estimation and its applications in a variety of scenarios [12, 14, 22, 32, 42]. We propose a novel gaze completion task and show its application to the problem of unsupervised mistake detection in egocentric video. Differently from previous works, we define and tackle the novel task of gaze completion, with the aim to reduce the uncertanty associated with gaze prediction.

Use of gaze in egocentric vision While many previous works focused on gaze estimation from video, few works investigated the use of gaze, estimated through a dedicated gaze tracker, as an input to support downstream egocentric vision applications. Specifically, previous investigations focused on discovering object usage [3], detecting privacy-sensitive situations [38], finding attended objects [27], assisting large language models in classification tasks [16], enhancing visual tasks [44, 46], improving egocentric human motion prediction [45], and aiding natural language processing tasks [36]. We show the effectiveness of gaze in mistake detection. Our method compares gaze trajectories predicted from visual data with gaze estimated through a gaze tracker to identify mistakes when predictions deviates from the ground truth.

Mistake Detection in Egocentric Videos Mistakes naturally occur in human activities. The ability to automatically detect them from egocentric video can be beneficial for an AR assistant to offer support. Identifying mistakes usually entails modeling procedural knowledge [4, 9, 34], skill assessment [10], action segmentation [11] or detecting forgotten actions [37]. Notably, previous works tackled the task in a supervised fashion, training models to classify an action segment as "correct" or "mistake" in manually annotated instances [34, 42]. While this approach is feasible in a closed-world scenario, it requires 1) a definition of what a mistake is, depending on the domain (e.g., kitchens vs the assembly line), 2) significant amounts of manually labeled data, which is expensive and requires expert knowledge. In this work, we tackle an unsupervised mistake detection task, in which models observe unlabeled video at training time and are tasked to detect mistakes from video at test time. Our unsupervised scheme is possible through the analysis of gaze attention patterns, which provide a supervisory signal to create a joint video-gaze model of normal behavior.

Video Anomaly Detection Our research also relates to the problem of Video Anomaly Detection (VAD), which involves recognizing abnormal or anomalous events within videos [23, 39]. A line of video anomaly methods are based on one-class classification, in which models are trained on normal videos and aim to identify divergence from the norm at test time [6, 8, 24, 25, 39, 40, 43]. Notably, anomaly detection in egocentric vision remains under-explored [26]. Similar to video anomaly detection, we aim to detect mistakes by determining video segments which deviate from statistics observed at training time [8]. Differently from previous works in video anomaly detection, we ground our predictions in an egocentric gaze estimation model, which acts as a proxy for modeling normal human behavior, hence effectively achieving mistake prediction detection when anomalous behavior is observed. Moreover, we go beyond the one-class assumption and show that our method can also be used in unsupervised settings where unlabeled correct and mistake examples are included at training time.

3. Proposed Approach

3.1. Mistake Detection Problem Setup

The mistake detection task consists in highlighting those parts of the video in which the user is making a mistake during the execution of a given activity. In our setup, at each time-step t, a model Φ takes as input a video V observed up to time-step t, $V_{1:t}$, and a 2D gaze trajectory $T_{1:t}$ obtained with a gaze tracker, where the i-th element of the trajectory $T_i^{(x,y)}$ is a 2D gaze fixation in frame V_i. Given this input, the model has to return a score $s_t = \Phi(V_{1:t}, T_{1:t})$ indicating whether a mistake is happening at the current time t. In this context, high s_t scores indicate the occurrence of a mistake, while low s_t scores indicate a correct action. We can hence see the mistake detection problem as a classification task, in which timestep t is classified as a mistake if $s_t > \theta$, where θ is a chosen threshold. We follow previous literature on anomaly detection [8, 39] and evaluate methods in a threshold-independent fashion by reporting the Receiver Operating Characteristics Area Under the Curve (ROC-AUC), where we consider "mistake" as the positive class[1]. For completeness, we also report the best F_1 score achieved considering the different thresholds, as well as its related precision and recall values.

3.2. Proposed Approach

At each timestep t, we trim the input video $V_{1:t}$ and gaze trajectory $T_{1:t}$ to the last observed F frames, hence considering $V_{t-F:t}$ and $T_{t-F:t}$ as inputs to our mistake detection method. (Figure 1(a)). Our method relies on two main components: a gaze completion model (Figure 1(b)), and a scoring function (Figure 1(c)).

Gaze Completion Model Figure 2 illustrates the proposed gaze completion model. The model takes as input the video $V_{t-F:t}$ and the first half of the input gaze trajectory $T_{t-F:t-F/2}$ and predicts a gaze trajectory $\hat{T}_{t-F/2:t}$ aligned to the remaining part of the ground truth trajectory $T_{t-F/2:t}$. The goal of this model is to predict where the user is looking in the video, conditioned on the initial trajectory. As we show in the experiments, the conditioning allows to reduce the uncertainty on gaze predictions and give a prior into the intention and characteristics of the user. For instance, the model can notice that the user is a novice from the partial input trajectory or get an understanding of the performed activity and adapt its prediction accordingly. We build on [19] and propose an encoder-decoder transformer-based architecture, including two approaches to condition gaze prediction on the input partial trajectory: channel fusion and correlation fusion.

Model Overview The input gaze trajectory $T_{t-F:t}$ is encoded into a stack of heatmaps Q obtained by centering a Gaussian distribution of standard deviation σ around the gaze points. The first half of the stack $Q_{1:F/2}$, corresponding to the input half trajectory $T_{t-F:t-F/2}$, is forwarded to the two trajectory fusion models (paths (1) and (2) in Figure 2), which inject information on the input trajectory at different semantic levels in the model. Input frames $V_{t-F:t}$ are processed by a token embedding layer which maps them to visual tokens with a convolution as in [19]. This mod-

[1] A true positive is a mistake correctly classified as a mistake, a true negative is a correct execution correctly classified as a correct execution, a false positive is a correct execution wrongly classified as a mistake, and a false negative is a mistake wrongly classified as a correct execution.

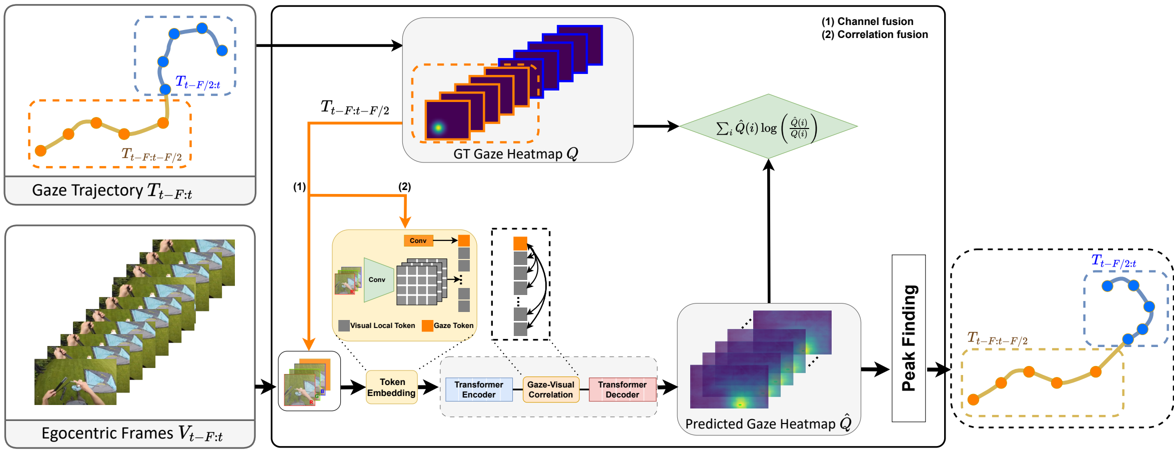


Figure 2. The model takes as input F RGB frames $V_{t-F:t}$ and a partial 2D gaze trajectory $T_{t-F:t-F/2}$ on the first $F/2$ input frames, and outputs a predicted trajectory $\hat{T}_{t-F/2:t}$ from the input video, conditioned on the input trajectory. The input trajectory is encoded as a spatio-temporal heatmap Q. Trajectory and RGB inputs are fused using two strategies, channel fusion, which adds gaze heatmaps as a separate channel (1), and correlation fusion, which uses a dedicated gaze-visual correlation module (2). We follow the design of [19] and process our inputs with a transformer encoder-decoder architecture which outputs a predicted gaze heatmap $\hat{Q}$, supervised via a standard Kullback–Leibler divergence loss using the ground truth unobserved trajectory $T_{t-F/2:t}$. The output trajectory $\hat{T}_{t-F/2:t}$ is recovered from $\hat{Q}$ using a peak finding operation.

ule is also responsible for mapping the input heatmaps to a single trajectory token. Input tokens are then processed by the transformer encoder, by the gaze-visual correlation module and finally by the transformer decoder to output a likely completion of the gaze in the form of heatmaps $\hat{Q}$, which are supervised with a standard Kullback–Leibler loss. Specifically, $\hat{Q}$ contains F heatmaps related to the F input frames $V_{t-F:t}$. The final trajectory $\hat{T}$ is obtained by finding the global maxima of the predicted gaze heatmaps.

Channel Fusion The channel fusion module (Figure 2(1)) adds the heatmaps in $Q_{1:F/2}$ to the first $F/2$ frames of the input video, $V_{t-F:t-F/2}$ as an additional channel. The values of this channel are set to zero for the remaining frames. This form of early fusion acts as a *soft conditioning* aiming to include information about the input gaze trajectory in the computation. Note that, in order to incorporate information about the relationships between haze and input frames, the model needs to learn how to compute suitable gaze representations from the additional input channel during training.

Correlation Fusion This approach (Figure 2(2)) aims to fuse visual tokens with the gaze trajectory token computed by the token embedding layer. Inspired by global-local fusion originally proposed in [19], this is done in two stages. First, within the transformer encoder, where attention between all visual features and the gaze token is computed. In this stage, correlations between all visual tokens and across visual and gaze tokens are leveraged to obtain a strong representation. Second, within the dedicated gaze-visual correlation module. Here, attention is computed only between the gaze token and the visual tokens, thus learning a dedicated attention mechanism which explicitly enriches the representation of each visual token with gaze information. Note that this fusion mechanism operates both at the early (through the encoder) and mid (through the gaze-visual correlation module) levels, thus allowing to leverage low-level (co-occurrences of gaze and visual features) and more semantic (co-occurrences of gaze and semantic visual concepts) information.

Scoring Function Our approach predicts the mistake confidence score s_t by comparing the predicted trajectory $\hat{T}_{t-F/2:t}$ with the ground truth one $T_{t-F/2:t}$ which is obtained by the gaze tracker of the wearable device. We explore four different ways to compare the two trajectories: Euclidean distance, Dynamic Time Warping (DTW), Heatmap, and Entropy.

Euclidean Distance This method consists in accumulating the Euclidean distances computed between corresponding points in each trajectory as the score s_t:

$$s_t = \sum_{i=F/2+1}^{F} \|T_{t-F+i} - \hat{T}_{t-F+i}\|. \tag{1}$$

Dynamic Time Warping This scoring function uses Dynamic Time Warping (DTW)[31] to measure the distance between trajectories:

$$s_t = \mathtt{DTW}(T, \hat{T}) \tag{2}$$

Where $\mathtt{DTW}$ returns the cost of aligning T to $\hat{T}$ according to the DTW algorithm[2].

Heatmap Differently from previous functions, this approach explicitly considers the probability values predicted by the model at each location.Specifically, we evaluate the likelihood of a ground truth eye fixation T_i obtained by the device under the predicted heatmap $\hat{Q}_i$, which can be computed as:

$$P(T_i|\hat{Q}) = \hat{Q}_i(T_i^x, T_i^y) \tag{3}$$

where T_i^x and T_i^y are the coordinates of the trajectory point $T_i = [T_i^x, T_i^y]$. The score associated to the predicted trajectory $\hat{T}$ is computed as the sum of the likelihoods of each trajectory point, considering the predicted heatmap $\hat{Q}$:

$$s_t = \sum_{i=F/2+1}^{F} P(T_{t-F+i}|\hat{Q}). \tag{4}$$

Entropy This is the only method which does not require ground truth gaze for computation. We consider this measure as a way to check whether mistakes are systematically characterized by uncertain gaze predictions. In this case, the score s_t is set as the mean entropy of all predicted heatmaps for a given trajectory $\hat{T}$. The entropy H of a single heatmap $\hat{Q}$ is given by:

$$s_t = -\frac{1}{F/2} \sum_{i=F/2+1}^{F} \sum_{x,y} \hat{Q}_{t-F+i}^{(x,y)} \log_2(\hat{Q}_{t-F+i}^{(x,y)}) \tag{5}$$

$Q_j^{(x,y)}$ is the value at coordinates (x, y) of heatmap Q_j.

4. Experiments and Results

4.1. Datasets and Implementation Details

We perform our experiments on three popular datasets.

EPIC-Tent [14] includes 7 hours of egocentric video of 29 subjects wearing a head-mounded GoPro and an SMI eye tracker while assembling a camping tent. The dataset includes egocentric video, gaze and labels indicating video segments in which users make mistakes. Subjects also rated their level of confidence in action execution in each clip in the videos. EPIC-Tent contains $151,689$ mistake frames and $384,558$ frames of correct executions, hence with a 28:72 ratio between correct and mistake frames. Since no official train-test split is available, we randomly split videos in training, validation and test sets roughly following a 60:15:25 ratio, obtaining $86,099$, $27,613$, and $37,977$ mistake frames in the training, validation, and test sets respectively. Since the dataset contains a single video per subject, there is no subject overlap between the three sets.

IndustReal [32] is designed for studying procedural tasks in industrial-like environments and consists of distinct training, validation, and test sets. The training set comprises $78,902$ frames, with 95.68% frames labeled as correct and 4.32% labeled as mistakes. The validation set includes $38,036$ frames, with 95.18% correct and 4.82% mistaken frames. The test set contains $90,105$ frames, with 92.53% correct and 7.47% mistaken frames.

HoloAssist [42] focuses on a variety of scenarios in which users perform tasks with the assistance of an expert. The training set comprises $11,614,033$ frames, with 94% frames labeled as correct and 6% labeled as mistakes. The test set contains $1,699,562$ frames, with 95% correct and 5% mistake frames.

Implementation Details We process input frames with a stride of 1 and set the batch size to 4 clips of 8 frames each. Weight decay is set to 0.07 to prevent overfitting. See the supplementary material for more details.

4.2. Supervision Levels and Compared Approaches

We compare the proposed approach to methods belonging to three different supervision levels: fully supervised, one-class classification, and unsupervised. All baselines described below are compared with a random baseline which assigns a random score to each input clip.

Fully Supervised Methods are trained assuming the availability of mistake labels for all image frames. We consider this class of methods to provide an upper-bound to performance when assessing one-class and unsupervised methods. We consider two approaches in this class: a TimeSformer [2] action recognition model which classifies the input video clips without access to any temporal context from actions executed before or after the current one, and a C2F [35] temporal action segmentation model operating on DINOv2 [28] features, which naturally performs action segmentation taking into account the temporal context in which an action (or mistake) is executed.

One-Class Classification Methods are trained only on *videos of correct executions*, following the standard setup of anomaly detection [8, 39]. In this context, we assume that the data is verified by an expert for correctness before being used for training. Note that this check does not require marking the temporal occurrence of mistakes, but only discarding any video which contains mistakes. For this class, we compare our method with respect to TrajREC [39] and MoCoDAD [8], two popular approaches for video anomaly detection based on the processing of human skeletal data. Since full human skeletons are not visible in egocentric videos, we replace skeletal data with hand joint keypoints[3]. We also adapt TrajREC and MoCoDAD to take a single gaze point instead of, or in addition to, the hand keypoints to assess the ability of such methods to leverage

[2]We used this implementation: https://pypi.org/project/fastdtw/.

[3]We use ground truth hand keypoints in HoloAssist and IndustReal, while we extract keypoints with https://github.com/open-mmlab/mmpose in EPIC-Tent.

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Table 1. Ablation of various scoring functions and fusion strategies on EPIC-Tent in the unsupervised setting. Best results perblock are underlined, while best global results are **in bold**. CH: channel fusion, CORR: correlation fusion.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

Table 2. Mistake detection results on EPIC-Tent. Best results are **in bold**, second best results are underlined.

eye-gaze information[4]. We compare these models to an instantiation of the proposed approach in which the gaze completion model is trained only on correct executions, hence effectively replicating a one-class scheme. We also compare with respect to a baseline which replaces the proposed gaze completion module with a simple gaze prediction component based on GLC [19]. Following the one-class setup, we train the GLC method of [19] on correct executions only and compare the predicted and ground truth gaze using the considered scoring functions.

Unsupervised Methods assume no knowledge of which examples are correct executions and which are mistakes. Hence, models are trained on a natural mix of correct and incorrect action executions. This is the least constrained case in which the collected data is not verified by an expert prior to training. We compare our model with TrajREC and MoCoDAD adapted as discussed above and with the gaze-prediction baseline GLC [19].

4.3. Performance of Proposed Model and Ablations

Table 1 reports the performance of the proposed approach in unsupervised settings, evaluating the considered scoring functions and gaze-video fusion strategies in the unsupervised mistake detection settings on EPIC-Tent. Rows 2-5 compare scoring functions when both fusion strategies are turned off and the model is not conditioned on previous trajectories. The entropy scoring function achieved an AUC of 0.51 and an F1 score of 0.41 (row 2), only marginally above the random baseline (F1 of 0.36), suggesting that high entropy in the predictions marginally correlates with the presence of a mistake. Alternative scoring functions improved the results, yielding an AUC of 0.55 and 0.56 when using Euclidean distance and DTW scoring functions, respectively (rows 3-4). The heatmap-based scoring function produced the best results, with an AUC of 0.57 and an F1 score of 0.45, significantly above random level (compare with the F1 score of 0.36 in row 1). The advantage of the heatmap scoring function is likely due to the better exploitation of the probability values computed by the gaze prediction model, as compared to other scoring functions. Hence, we adopted the heatmap-based scoring as our primary method in following comparisons.

Rows 6-7 compare approaches using one of the two fusion strategies. While both fusion strategies improve results (compare rows 6-7 with 5), the proposed correlation strategy (CORR) systematically outperforms channel fusion, obtaining an AUC score of 0.65 and an F1 score of 0.50 and doubling the recall of the random baseline (0.82 vs 0.42) with better precision (0.36 vs 0.29). Combining the two fusion strategies (row 8) leads to an AUC of 0.69 and an F1 score of 0.51 (*+0.12* and *+0.06* compared to the standard heatmap method - row 5). This configuration is the one referred to as “ours” in future comparisons.[5]

4.4. Comparison with the state of the art

EPIC-Tent Table 2 compares the proposed mistake detection approach with competitors on EPIC-Tent, according to the three considered levels of supervision. C2F outperforms TimeSformer in all evaluated metrics, particularly in the F1 score (0.58 of C2F vs 0.49 of TimeSformer and 0.36 of the random baseline), suggesting that C2F is more adept at capturing temporal reasoning, which is crucial for identifying mistakes in dynamic activities. However, the reliance on fully labeled datasets poses limitations for both methods. One-class methods for anomaly detection adapted to take only gaze as input, namely, TrajREC (G) and MoCoDAD (G), show minor improvements over the random baseline in terms of AUC score ($0.50 - 0.51$ vs 0.51) and only small

[4]See supplementary material for more details.

[5]See supplementary material for additional ablations.

improvements in F1 score ($0.40-0.43$ vs 0.36). High recall values, paired with low precision, suggest that these methods tend to classify most clips as mistakes. Incorporating hand skeleton data instead of gaze, namely, TrajREC (H) and MoCoDAD (H), leads to slight improvements, as evidenced by F1 scores of 0.44 and 0.45, and AUC values of 0.55 and 0.60. Combining (H) and (G) models through late fusion (denoted with *) improves the results, with MoCoDAD (H+G)* achieving an F1 score of 0.49 and an AUC of 0.65, suggesting that the signals captured by gaze-based and hand-based analyses are complementary. The one-class approach based on GLC [19] gaze prediction shows improved results compared to previous one-class methods, being only slightly less effective than the late fused method, despite not analyzing any hand-based information. This highlights the value of leveraging gaze analysis for mistake detection. "'

Finally, the proposed method based on gaze completion obtains the best results, yielding an AUC of 0.69, an F1-score of 0.52, a precision of 0.37, and a recall of 0.85, which amount to relative improvements of $+5/13\%$[6] with respect to the best approach GLC and $+35/44\%$ with respect to the random baseline. Late-fusing our approach with MoCoDAD (H) achieves enhanced results with an AUC of 0.72 and an F1 score of 0.54, improving over compared approaches, suggesting that gaze analysis can further benefit from integration with approaches based on different cues. It is worth noting that our best method remarkably achieves the same AUC score of 0.72 as the best supervised approach and a comparable $F1$ score (0.54 vs 0.58) without access to labels during training. We compare unsupervised approaches in the bottom part of Table 2. Similarly to the one-class case, TrajREC (H) and MoCoDAD (H) slightly improve over the random baseline (e.g., 0.40 and 0.41 vs 0.36 of the random baseline in F1 score). GLC outperforms these former two methods obtaining an F1 score of 0.44 and an AUC score of 0.61, which are lower than the scores of 0.46 and 0.66 obtained in the one-class setting. In these settings, the proposed method achieves an F1 score of 0.51, which is comparable to the score obtained in one-class settings 0.52 with similar and recall values and the same AUC of 0.69, despite the unsupervised setting being more challenging. Late fusion with MoCoDAD (H) brings some additional improvements, with an F1 score of 0.52 and AUC of 0.70.

Table 3 compares the proposed method with competitors on HoloAssist, which presents a more varied and expansive context than EPIC-Tent, making mistake detection more challenging. The random baseline achieves an F1 score of only 0.04. One-class methods like TrajREC and MoCoDAD improve over the baseline but tend to classify most clips as mistakes, with high recall values (0.96 and 0.94). Hand keypoint-based methods show improvements, with

[6]We compute the relative improvement of b with respect to a as $\frac{b-a}{b}$

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.04	0.02	0.39	0.50
TimeSformer [2]	Fully Supervised	0.21	0.35	0.13	0.58
C2F [35]	Fully Supervised	**0.38**	**0.37**	**0.40**	**0.65**
TrajREC (G) [39]	One-Class	0.09	0.04	**0.96**	0.50
MoCoDAD (G) [8]	One-Class	0.11	0.06	0.94	0.51
TrajREC (H) [39]	One-Class	0.19	0.11	0.72	0.56
MoCoDAD (H) [8]	One-Class	0.17	0.10	0.71	0.55
TrajREC (H+G) [39]	One-Class	0.13	0.07	0.68	0.52
MoCoDAD (H+G) [8]	One-Class	0.14	0.08	0.62	0.52
TrajREC (H+G)* [39]	One-Class	0.20	0.12	0.71	0.56
MoCoDAD (H+G)* [8]	One-Class	0.21	0.12	0.75	0.57
GLC [19]	One-Class	0.19	0.11	0.56	0.60
Ours	One-Class	0.22	0.14	0.59	0.61
Ours + MoCoDAD (H)*	One-Class	**0.26**	**0.16**	0.73	**0.63**
TrajREC (G) [39]	Unsupervised	0.05	0.03	**0.92**	0.50
MoCoDAD (G) [8]	Unsupervised	0.07	0.04	**0.92**	0.50
TrajREC (H) [39]	Unsupervised	0.11	0.07	0.32	0.56
MoCoDAD (H) [8]	Unsupervised	0.14	0.10	0.25	0.55
MoCoDAD (H+G)* [8]	Unsupervised	0.15	0.11	0.25	0.56
GLC [19]	Unsupervised	0.10	0.06	0.34	0.54
Ours	Unsupervised	0.18	0.12	0.40	0.59
Ours + MoCoDAD (H)*	Unsupervised	**0.21**	**0.15**	0.40	**0.60**

* Late fusion

Table 3. Mistake detection result on HoloAssist.

TrajREC (H) slightly outperforming MoCoDAD (H). GLC demonstrates better AUC and F1 scores, with more balanced precision and recall metrics. Our approach achieves an AUC of 0.61 and an F1 score of 0.22 in one-class settings, with the best results from late fusion with MoCoDAD, yielding an F1 score of 0.26 and an AUC of 0.63. In the unsupervised scenario, TrajREC (G) and MoCoDAD (G) show limited effectiveness, while the (H) approaches perform slightly better. Our method achieves an AUC of 0.59 and an F1 score of 0.18, showing robustness across evaluation settings and relative improvements over GLC, with gains of $+9\%$ and $+80\%$, respectively. Combining with MoCoDAD (H) further enhances performance.

Results on IndustReal, shown in Table 4, confirm the trends observed in HoloAssist. TrajREC and MoCoDAD bring small improvements over the random baseline in H, G, and H+G configurations. Our method outperforms competitors with improvements over GLC of $+5\%$ in AUC and $+14\%$ in F1 in one-class settings, and $+6\%$ in AUC in unsupervised settings, while late fusion with MoCoDAD (H) does not improve performance due to MoCoDAD's reduced effectiveness in this scenario.

4.5. Contribution of gaze across scenarios

In this section, we analyze the performance of our method with respect to scenarios in order to assess under which conditions gaze analysis is more or less predictive of mistakes. **Action Complexity** We investigated how action complexity correlates with gaze-predicted mistakes in procedural tasks. We asked 40 volunteers to rate the complexity of performing actions contained in HoloAssist without looking (1 = easy, 5 = difficult). We then compared the complexity of the action associated to a given video segment to the ability

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.12	0.06	**0.62**	0.51
TimeSformer [2]	Fully Supervised	0.20	0.12	0.35	0.58
C2F [35]	Fully Supervised	**0.31**	**0.29**	0.31	**0.67**
TrajRE(G) [39]	One-Class	0.17	0.09	0.90	0.53
MoCoDAD(G) [8]	One-Class	0.18	0.10	**0.91**	0.55
TrajREC(H) [39]	One-Class	0.21	0.12	0.88	0.57
MoCoDAD(H) [8]	One-Class	0.22	0.13	0.81	0.60
TrajREC(H+G) [39]	One-Class	0.18	0.10	0.86	0.55
MoCoDAD(H+G) [8]	One-Class	0.19	0.11	0.79	0.58
TrajREC(H+G)* [39]	One-Class	0.21	0.12	0.88	0.58
MoCoDAD(H+G)* [8]	One-Class	0.22	0.13	0.82	0.61
GLC [19]	One-Class	0.21	0.15	0.33	0.60
Ours	One-Class	0.24	**0.18**	0.35	0.63
Ours + MoCoDAD (H)*	One-Class	**0.26**	0.17	0.60	**0.65**
TrajREC (G) [39]	Unsupervised	0.11	0.06	**0.92**	0.51
MoCoDAD (G) [8]	Unsupervised	0.11	0.06	**0.92**	0.51
TrajREC (H) [39]	Unsupervised	0.15	0.11	0.28	0.55
MoCoDAD (H) [8]	Unsupervised	0.16	0.12	0.29	0.57
MoCoDAD (H+G)* [8]	Unsupervised	0.17	0.12	0.30	0.57
GLC [19]	Unsupervised	**0.21**	0.15	0.33	0.58
Ours	Unsupervised	**0.21**	**0.16**	0.33	**0.62**
Ours + MoCoDAD (H)*	Unsupervised	0.20	0.15	0.32	0.61

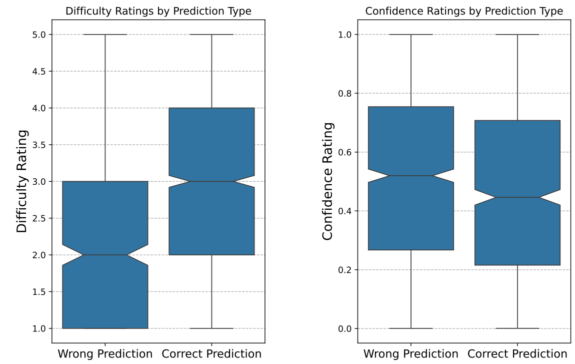
* Late fusion

Table 4. Mistake detection result on IndustReal.

of our model to make a correct prediction (which we term "success"). Results (see Figure 3a) showed a positive correlation between difficulty and prediction success, measured with a Point Biserial Correlation of 0.3843, with $p < 0.05$[7]. This suggests that our method is particularly effective in the case of complex actions which cannot be carried out without looking, while less effective in the case of trivial tasks.

Confidence Level On the EPIC-Tent dataset, we compared if the self-rated confidence score reported by camera wearers was correlated to the success of our method. Results (see Figure 3b) obtained a Point Biserial Correlation of -0.1137, $p < 0.05$ indicating a small but significant negative correlation: our method is most effective when the self-rated confidence is higher. This suggests that gaze-based analysis is more effective in the case of novices, which reported lower confidence and probably rely more on visual observations when executing their tasks.

Action Type We finally assess whether the type of the performed action affects the performance of our method. To this aim, we grouped actions contained in all three datasets in four categories (Hand-Eye Coordination, Object Manipulation, Task Preparation, Inspection/Verification)[8]. We hence computed the number of co-ocurrences between success or failure of our method and the different action classes. Results (see Figure 4) show a Cramer's V statistic of 0.27 (a moderate correlation of 0.27 in a $0-1$ scale) with a p-value $p < 0.05$. Gaze-based analysis proves particularly useful in the case of actions requiring hand-eye coordination and object manipulation abilities, while less effective



(a) (b)

Figure 3. Distributions of difficulty ratings (a) and execution confidence ratings (b) with respect to wrong and correct predictions.

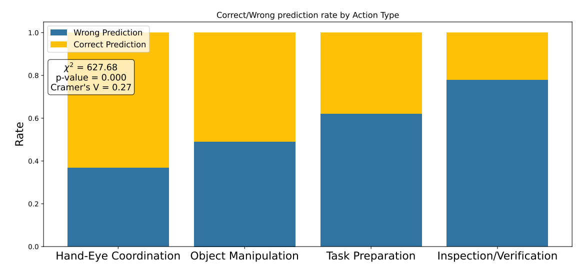


Figure 4. Distributions of Correct/Wrong pred. by action type.

for generic actions such as task preparation and inspection.[9]

5. Conclusion

We proposed to perform mistake detection in egocentric videos in an unsupervised way, leveraging gaze signals. We introduced a novel *gaze completion task*, where gaze trajectories are predicted based on observed video and partial gaze data, and an approach to tackle this task. Mistake detection is performed comparing predicted trajectories with ground truth, identifying instances where gaze becomes unpredictable as potential mistakes. Experimental validation on EPIC-Tent, HoloAssist, and IndustReal demonstrates the efficacy of our method, surpassing traditional one-class techniques and other unsupervised mistake detection methods. Our method is ranked first on the HoloAssist Mistake Detection challenge. Code will be publicly shared.

Acknowledgments

This research has been supported by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006.

[7] We use Point Biserial Correlation as action difficulty is a continuous variable while the success of our method is a binary one.

[8] See supplementary material for more details.

[9] See supp. for more details and per class F1 and AUC scores. Qualitative results are reported in the supplementary material

References

[1] Mohammad Al-Naser, Shoaib Ahmed Siddiqui, Hiroki Ohashi, Sheraz Ahmed, Nakamura Katsuyki, Sato Takuto, and Andreas Dengel. Ogaze: Gaze prediction in egocentric videos for attentional object selection. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2019. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 5, 6, 7, 8

[3] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *British Machine Vision Conference*, 2014. 2

[4] Guodong Ding, Fadime Sener, Shugao Ma, and Angela Yao. Every mistake counts in assembly. *ArXiv*, abs/2307.16453, 2023. 3

[5] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 314–327. Springer, 2012. 2

[6] Jianfeng Feng, Fa-Ting Hong, and Weishi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14004–14013, 2021. 3

[7] Shijia Feng, Michael Wray, Brian Sullivan, Youngkyoon Jang, Casimir J. H. Ludwig, Iain Gilchrist, and Walterio W. Mayol-Cuevas. Are you struggling? dataset and baselines for struggle determination in assembly videos. *ArXiv*, abs/2402.11057, 2024. 1

[8] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10318–10329, 2023. 2, 3, 5, 6, 7, 8

[9] Alessandro Flaborea, Guido Maria D'Amely di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, and Fabio Galasso. Prego: online mistake detection in procedural egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18483–18492, 2024. 2, 3

[10] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, 2014. 3

[11] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, and Behzad Dariush. Weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10128–10138, 2023. 3

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. *ArXiv*, abs/1803.09125, 2018. 2

[14] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain D. Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4461–4469, 2019. 2, 5

[15] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain D. Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4461–4469, 2019. 2

[16] Robert Konrad, Nitish Padmanaban, J. Gabriel Buckmaster, Kevin C. Boyle, and Gordon Wetzstein. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear, 2024. 2

[17] Fatemeh Koochaki and Laleh Najafizadeh. Predicting intention through eye gaze patterns. *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4, 2018. 2

[18] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M. Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184, 2016. 2

[19] Bolin Lai, Miao Liu, Fiona Ryan, and James Rehg. In the eye of transformer: Global-local correlation for egocentric

gaze estimation. *British Machine Vision Conference*, 2022. 2, 3, 4, 6, 7, 8

[20] M. Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311–28, 1999. 2

[21] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 2

[22] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[23] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[24] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 3

[25] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8022–8031, 2023. 3

[26] Mana Masuda, Ryo Hachiuma, Ryo Fujii, and Hideo Saito. *Unsupervised Anomaly Detection of the First Person in Gait from an Egocentric Camera*, page 604–617. Springer-Verlag, Berlin, Heidelberg, 2020. 3

[27] Michele Mazzamuto, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Learning to detect attended objects in cultural sites with gaze signals and weak object supervision. *J. Comput. Cult. Herit.*, 2024. 2

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[29] Candace E. Peacock, Ben Lafreniere, Ting Zhang, Stephanie Santosa, Hrvoje Benko, and Tanya R. Jonker. Gaze as an indicator of input recognition errors. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA), 2022. 2

[30] Jeff Pelz and Roxanne Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision research*, 41: 3587–96, 2001. 2

[31] Hiroaki Sakoe. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:159–165, 1978. 4

[32] Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons van der Sommen, et al. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4365–4374, 2024. 2, 5

[33] Luigi Seminara, Giovanni Maria Farinella, and Antonino Furnari. Differentiable task graph learning: Procedural activity representation and online mistake detection from egocentric videos. In *Advances in Neural Information Processing Systems*, 2024. 2

[34] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 1, 3

[35] Dipika Singhania, Rahul Rahaman, and Angela Yao. C2f-tcn: A framework for semi-and fully-supervised temporal action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11484–11501, 2023. 5, 6, 7, 8

[36] Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, 2020. 2

[37] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4669–4677, 2015. 3

[38] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pages 1–10, 2019. 2

[39] Alexandros Stergiou, Brent De Weerdt, and Nikos Deligiannis. Holistic representation learning for multitask trajectory anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2, 3, 5, 6, 7, 8

[40] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[41] Hamed Rezazadegan Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282, 2019. 2

[42] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20270–20281, 2023. 1, 2, 3, 5

[43] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. In *British Machine Vision Conference*, 2015. 3

[44] Yushi Yao, Chang Ye, Junfeng He, and Gamaleldin F. Elsayed. Teacher-generated spatial-attention labels boost robustness and accuracy of contrastive models. In *2023*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23282–23291, 2023. 2

[45] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas J. Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, 2022. 2

[46] Yuchen Zhou, Linkai Liu, and Chao Gou. Learning from observer gaze:zero-shot attention prediction oriented by human-object interaction recognition, 2024. 2