Team PKU-WICT-MIPL Ego4D Look At Me Challenge 2023 Technical Report

Xiyu Wei[1] Dejie Yang[1] Yuxin Peng[1] Yang Liu[1,2]*
[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Key Laboratory of General Artificial Intelligence, BIGAI
{2000013085,ydj}@stu.pku.edu.cn {pengyuxin,yangliu}@pku.edu.cn

Abstract

This report presents the solutions developed by our team, 'PKU-WICT-MIPL,' for the Ego4D Looking At Me (LAM) Challenge at CVPR 2023. The objective of the LAM Challenge is to determine whether a person, whose visible face has been localized in an egocentric video clip, is looking at the camera wearer. However, egocentric videos suffer from blurry face regions and large frame quality variance, which limits models to recognize eye contact for LAM detection. To address this issue, we propose GazePose, which exploits priors from pose-related information to improve eye contact recognition for LAM detection. Our proposed model ranked 1st in the Ego4D Looking At Me Challenge of CVPR 2023 surpassing the second-best approach by a margin of 2% in mean average precision (mAP).

1. Introduction

Understanding human eye interaction is crucial for developing more capable virtual assistants and social robots. The egocentric videos offer a unique view and important cues for recognizing mutual eye interaction. Specifically, the egocentric *looking at me (LAM)* task focuses on classifying whether someone is looking at the camera-wearer. Progress in egocentric eye contact understanding could significantly enhance the capabilities of these technologies.

Nevertheless, the task of detecting eye contact in egocentric videos poses a challenge: the rapid movement of wearable cameras often leads to a decrease in image quality within the videos. Consequently, the blurry and ambiguous representation limits model to obtain fine-grained representations of the eyes from video frames. To further quantify this effect, we used the Laplacian operator to evaluate image quality and divided the validation set of Ego4D [4] into two level splits according to the image quality: low, high. The conventional LAM model [4] achieves an average precision (AP) of 76.83% on low-quality images compared to 93.62% on high-quality images. These results highlight the importance of addressing image quality issues when designing algorithms for eye contact detection in egocentric videos.

To address the problems, we propose a novel framework, named GazePose, which integrates pose-related priors to improve eye contact detection, especially in low-quality frames. Estimating the head pose and facial landmark, i.e., the orientation of head and facial expression, can help to confirm whether a person is making eye contact with the camera-wearer. The underlying motivation for focusing on the broader context of head and face regions, rather than solely relying on the eye region, stems from the fact that these regions offer greater robustness, especially in scenarios where image quality is compromised. Specifically, GazePose consists of two main components. Firstly, to refine the spatial representation of the single frame, GazePose incorporates a Transformer decoder that aggregates the feature map with head pose and facial landmarks. The cross-attention captures the correlation between visual feature maps and pose-related priors, leading to improved key region features for eye contact detection. Secondly, GazePose introduces a temporal dynamic refinement approach for high-quality frame selection. This method deals with the inconsistency in image quality across the temporal axis by assuming that the distribution of image quality variance across frames follows a Gaussian distribution. By utilizing this assumption, we can selectively remove frames with low-quality images while preserving frames with high-quality images for subsequent processing. Overall, GazePose can capture both spatial and temporal cues which results in reliable eye contact detection in egocentric videos.

2. Methodology

2.1. Problem Formulation.

Given an egocentric video clip consisting of $2N + 1$ frames $\{\boldsymbol{I}_i\}_{i=t-N}^{t+N}$, each frame $\boldsymbol{I}_i \in \mathbb{R}^{H_0 \times W_0 \times 3}$ represents a cropped image of the face of the same person, where $H_0 \times W_0$ is the resolution of each frame. The goal is to

*Corresponding author

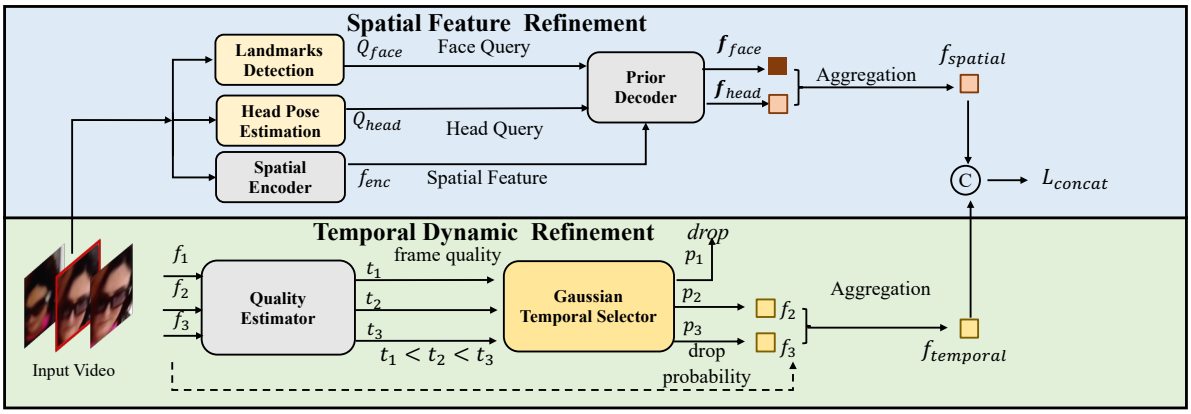

Figure 1. **The overall framework of our proposed *GazePose*.** To explore the priors of eye contact to improve LAM detection, we propose a Transformer-based framework to aggregate face representation, 3D head pose and 2D facial landmarks to assist LAM detection with a temporal dynamic refinement module to select high-quality frames and enhance temporal feature for LAM.

determine whether the person in the video clip has eye contact with the camera wearer, indicated by a binary label $y \in \{0, 1\}$, where 0 means no eye contact and 1 indicates the presence of eye contact.

2.2. GazePose.

Figure 1 illustrates the architecture of our GazePose. The GazePose consists of Spatial Feature Refinement module and Temporal Dynamic Refinement module.

Spatial Feature Refinement: The pose-related information such as head pose and facial landmarks can provide priors when the eye region is blurred. Thus, to improve the accuracy of eye contact detection in low-quality frames, we introduce head pose and facial landmarks as cues. Firstly, to explore the priors of head pose and facial landmarks, we extract the head pose features and facial features from the off-the-shelf head pose estimation [6] and facial landmarks detection[1] models respectively. To balance performance and efficiency, we only extract the priors on the central frame. Moreover, to obtain the central frame representation, we adopt a visual backbone to obtain the feature map $\boldsymbol{f}_c$. Secondly, to further explore the spatial correlation of visual features, we introduce a prior decoder to model the relations between the cues ($\boldsymbol{f}_{head}$ and $\boldsymbol{f}_{face}$) and frame representation. Specifically, we use a spatial transformer encoder to extract spatial features $\boldsymbol{f}_{enc} \in \mathbb{R}^{L \times D}$ from feature map patches of central frame $\boldsymbol{f}_c$, where L is the number of patches, and D is the hidden space dimension. Then, we take the cues $\boldsymbol{f}_{head}$ and $\boldsymbol{f}_{face}$ as the queries of and apply the cross attention mechanism between head queries $\boldsymbol{Q}_{head}$, landmark queries $\boldsymbol{Q}_{face}$ and frame spatial features $\boldsymbol{f}_{enc}$ to decode the eye contact related features $\boldsymbol{f}_{head}$ and $\boldsymbol{f}_{face}$ respectively.

$$[\boldsymbol{f}_{head}, \boldsymbol{f}_{face}] = \mathrm{D}_e(\boldsymbol{Q}, \boldsymbol{f}_{enc}) \in \mathbb{R}^{(N_h+N_l)\times D} \quad (1)$$

where $\boldsymbol{Q} = [\boldsymbol{Q}_{head}, \boldsymbol{Q}_{face}]$, D_e is the transformer decoder,$\boldsymbol{f}_{head}$ and $\boldsymbol{f}_{face}$ are the outputs of the decoder. The $\boldsymbol{f}_{enc}$ will serve as the key and value in the cross attention. N_h and N_l are the number of head queries $\boldsymbol{Q}_{head}$ and landmark queries $\boldsymbol{Q}_{face}$. Finally, we concatenate the $\boldsymbol{f}_{head}$ and $\boldsymbol{f}_{face}$ and feed it into a MLP to produce the final feature as spatial enhanced feature $\boldsymbol{f}_{spatial} \in \mathbb{R}^D$ which is the output of this branch and will be used to identify the eye contact in the final classifier

Temporal Dynamic Refinement: The image quality across frames in egocentric videos can vary, and low-quality frames may disturb the prediction of the model. Thus, we propose a temporal dynamic refinement branch to selectively remove frames with low-quality images : Firstly, we introduce Laplacian variance to estimate the quality t_i of each video frames $\{\boldsymbol{I}_i\}_{i=t-N}^{t+N}$, where the higher the Laplacian variance, the better the image quality [1]. And we propose a Gaussian Temporal Selector to select the high-quality frames and drop the low-quality for more accurate representation of the clip to select the frames with sufficient visual information according to the image quality score (Laplacian variance). Secondly, to aggregate the temporal information of the selected features, we encode them by a bidirectional LSTM to obtain the temporal enhanced feature $\boldsymbol{f}_{temporal} \in \mathbb{R}^D$.

$$\boldsymbol{f}_{temporal} = LSTM([\boldsymbol{I}_{k1}, \boldsymbol{I}_{k2}, ..., \boldsymbol{I}_{kn}]) \quad (2)$$

where $[k1, k2, ..., kn]$ is the selected index value. Finally, we concatenate the $\boldsymbol{f}_{temporal}$ and $\boldsymbol{f}_{spatial}$ (from spatial feature refinement module) as the final representation of the clip for eye contact. We use a MLP as classifier to identify the eye contact and apply a contact loss $L_{contact}$ to optimize the whole network.

3. Experiments

Dataset & Metrics We experiment on the Ego4D LAM dataset, which consists of 572 video clips in total. The number of samples of train/val/test split is 389/50/133 respectively. We use mAP and accuracy as the evaluation metrics according the rules of the LAM challenge[4].

Training Details We initialize the backbone and the Bi-LSTM from a pretrained Gaze360 [5] model, and use the

[1] https://github.com/google/mediapipe

pre-trained Hopenet[6] as the initial parameters of the head pose estimation module except the feed forward network (FFN). We use adam optimizer and set learning rate as 0.0005 for models using baseline-architecture and 0.0001 for GazePose. We froze the head pose estimation and Landmarks Detection modules except for the FFN when training. We apply a cosine learning rate decay strategy.

Method	mAP	Acc
Baseline[4]	71.92%	85.98%
HHW	74.00%	85.00%
NPT	73.00%	93.00%
GazeTR [3]	$72.01 \pm 0.82\%$	$82.31 \pm 2.22\%$
Looking [2]	$73.11 \pm 1.45\%$	$84.61 \pm 1.84\%$
GazeHead[7]	$74.12 \pm 2.32\%$	$85.36 \pm 3.10\%$
GazePose(ours)	$\mathbf{77.02 \pm 0.78\%}$	$\mathbf{87.35 \pm 2.09\%}$
Ensemble(ours)	**78.80%**	**92.47%**

Table 1. **Performance of different methods on the test set.** The method named HHW and NPT is from the second-best and third-best approach on the leaderboard. For GazeTR, Looking, GazeHead and GazePose, we conduct five times repeated experiments.

Comparisons with other methods. The comparison results are shown in Table 1. We implement GazeTR [3] and Looking [2] on the Ego4D dataset for comparison. GazePose outperforms other methods on mAP and Acc. It is worth mentioning that although GazeHead can achieve mAP of 77.39% in our previous technical report, we find that the performance of GazeHead fluctuates significantly when the experiment is repeated several times due to the loss of processing of low quality frames. Due to the introduction of facial landmarks and Gaussian temporal selector, GazePose performs more stable, which obtains 2.9% and 1.9% improvement on mAP and accuracy. Finally, we use XGBoost library to ensemble several GazePose with different hyperparameters (batch size, learning rate etc.). And the final ensemble model can get 6.88% improvement on mAP compared with baseline.

Spatial Feature	Temporal Dynamic	mAP	Acc
×	×	71.92%	85.98%
✓	×	76.31%	87.02%
×	✓	73.85%	86.18%
✓	✓	**77.51%**	**88.11%**

Table 2. Ablation study on Spatial Feature Refinement and Temporal Dynamic Refinement on Ego4D [4].

Effectiveness of Each Module. As shown in Table 2, we implement an ablation study on each module of GazePose. We evaluate the effectiveness of Spatial Feature Refinement and Temporal Dynamic Refinement. According to the table, the model with only temporal dynamic refinement achieves 1.93% absolute gain which shows that the Gaussian temporal selector indeed brings improvement by highlighting the contribution of high-quality frames. Meanwhile, with only spatial feature refinement, the model can achieve 4.39% absolute gain. It reveals the fact that spatial feature refinement module plays an important role in our design by providing potential face structure cues in social interactions. All the results together indicate that both of the components are essential for accurate eye contact detection.

4. Limitations and Future work

Figure 2. Examples showing the images and our LAM detection results. The red demonstrates a failure case (incorrect prediction).

To better explore the performance of the model with different image quality and study the possible limitations, we evaluate our model on different quality split on the validation set of Ego4D [4]. In Figure 2, we present the correctly and mistakenly predicted by our GazePose. The performance of our model in recognizing videos with very poor quality may produce errors. These videos are also difficult for humans to distinguish, thus they are also very challenging for the model. In the future, we will enhance the LAM detection of these poor quality videos by using their context (the segments before and after the video segment).

References

[1] Raghav Bansal, Gaurav Raj, and Tanupriya Choudhury. Blur image detection using laplacian operator and open-cv. In *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, 2016. 2

[2] Younes Belkada, Lorenzo Bertoni, Romain Caristan, Taylor Mordan, and Alexandre Alahi. Do pedestrians pay attention? eye contact detection in the wild. *arXiv*, 2021. 3

[3] Yihua Cheng and Feng Lu. Gaze estimation using transformer. *arXiv*, 2021. 3

[4] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 2, 3

[5] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019. 2

[6] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPRW*, 2018. 2, 3

[7] Xiyu Wei, Dejie Yang, Minghang Zheng, Qingchao Chen, Yuxin Peng, and Yang Liu. Team pku-wict-mipl ego4d look at me challenge 2022 technical report. 2022. 3