# Exploring adaptation of VideoMAE for Audio-Visual Diarization Social @ Ego4d Looking at me Challenge

Yinan He

Beijing University of Posts and Telecommunications

`heyinan@bupt.cn`

Guo Chen

State Key Lab for Novel Software Technology, Nanjing University

`chenguo1177@gmail.com`

## Abstract

*In this report, we present the transferring pretrained video mask autoencoders(VideoMAE) to egocentric tasks for Ego4d Looking at me Challenge. VideoMAE is data-efficient pretraining model for self-supervised video pre-training and can easily transfer to downstream tasks. We show that the representation transferred from VideoMAE has good Spatio-temporal modeling and the ability to capture small actions. We only need to use egocentric data to train 10 epochs based on VideoMAE which pretrained by the ordinary videos acquired from a third person's view, and we can get better results than baseline on Ego4d Looking at me Challenge.*

## 1. Introduction

Ego4d [3] has organized and published a large-scale, widely distributed egocentric video dataset. It makes many first-person perceptual tasks possible and promotes the development of related fields.

In human communication, "look at me" is an indispensable action to interact with others. In 2022Ego4D challenge@ECCV, we were asked to identify the communicative acts towards the camera-wearer. Specifically, given a video, the face of the social partner has been localized and identified, and each visible face should be classified as to whether is looking at the camera-wearer or not.

To alleviate this challenge, we take a face sequence as our model input and get the probability of whether this face in the clip looks at the camera-wearer. We replace the LSTM model with VideoMAE [8] on the official baseline, and simply fine-tune the videoMAE pretrained on Something-Something-V2 [2] with Ego4d annotations. *The code is available at https://github.com/yinanhe/ego4d-look-at-me.*

## 2. Approach

Masked Autoencoders is a new unsupervised pre-training paradigm. It is to mask the input part and achieve the pre-training result by restoring and reconstructing the mask part. MAE [4] has been proved to be a strong competitor of pre-training methods widely used in computer vision. The pre-training model based on MAE has achieved very high fine-tuning accuracy on a wide range of vision tasks of different types and complexity. Video masked autoencoders(VideoMAE) [8] uses self-supervised video pre-training based on a video mask tube, and shows an efficient learning method under a small amount of data.

Look at me aims to classify the given face in the sequence. The sequence $V$ of face $F_n$ with $T$ frames sampled in stride $s$ is denoted as $V_{F_n} = \{I_t^{F_n} \ t \in [1, T]\}$, where $I$ is the face image cropped according to face region. We train the model by cropped face sequences $V$ and their annotations. We obtain feature representation by fine-tuning VideoMAE on the Ego4d dataset, to take advantage of its powerful transfer capability. Then features are input to a linear layer to predict 2 categories.

## 3. Experiments

### 3.1. Experimental Configuration

**Dataset.** For our experiments, we use the Ego4D [3] Look-at-me benchmark. The dataset for this challenge is composed of 38242 clips. In the dataset, the data that is not looking at me is ten times more than the data that is looking at me. We sample face sequences from clips with $stride = 13$ during training, and $stride = 1$ during test. We extract three frames before and after the current frame as the input of the sequence model in each step. The blank

Table 1. Fine-tuning settings of VideoMAE.

|  | VideoMAE-B | VideoMAE-L |
|---|---|---|
| Resolution | 224 | |
| Optimizer | AdamW [7] | |
| Momentum | $\beta_1, \beta_2 = 0.9, 0.999$ | |
| Weight decay | 0.05 | |
| Learning rate schedule | cosine [6] | |
| Start learning rate | 1e-6 | |
| End learning rate | 1e-6 | |
| Batch size | 512 | 256 |
| Learning rate | 5e-6 | 4e-6 |
| Warmup epoch | 3 | |
| Epochs | 10 | |
| Scale | [0.08, 1.00] | |
| Jitter aspect ratio | [0.75, 1.33] | |
| Color jitter | 0.4 | |
| Rand augment [1] | rand-m7-n4-mstd0.5-inc1 | |

Table 2. Results of Experiments.

|  | mAP | Acc |
|---|---|---|
| Baseline | 0.66 | 0.74 |
| VideoMAE-B | 0.69 | 0.91 |
| VideoMAE-L | 0.73 | 0.93 |

frame will be used as a fill when the interval frame does not exist.

**Training details.** It is difficult to capture the changes in the eyes because they are very small. We use VideoMAE pretrained on Something-Something-V2 as our backbone. Finetuning settings are shown in Table. 1.

### 3.2. Results

In Table. 2, we evaluate our fine-tuning VideoMAE in the Ego4D Look at me benchmark and compare it with the challenge baseline: LSTM [5]. Our results show that VideoMAE-L outperforms the baseline by 7% mAP and 19% accuracy.

## 4. Conclusion

We implement VideoMAE as a pre-training model in a simple and efficient way for fine-tuning in egocentric video, which is used to perceive the interaction relationship of people in egocentric. Our methods have achieved competitive performance in the Look at me challenge. This method can make good use of the existing public video pre-training model, and easily transfer the representation learned in the third-person video to the egocentric perspective task.

**Limitaions.** Since we do not add more tricks to the training strategy, our method is vulnerable to the impact of data imbalance.

## References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 2

[2] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[6] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 2

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[8] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1