ACM MM Challenge 21 technique report

LIWEI JIN[*] and HAOYUE CHENG[*], Nanjing University and SenseTime Research, China

We use several open-source pre-trained models and ensemble them. Also, we use self-training method and OCR to get better results.

CCS Concepts: • **Large-Scale datasets pretrained models**; • **visual transformer**; • **self-training**; • **OCR**;

ACM Reference Format:
Liwei Jin and Haoyue Cheng. 2021. ACM MM Challenge 21 technique report. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/1122445.1122456

According to the challenge rules, we use open-source models pre-trained on large-scale open-source datasets, they are CLIP visual transformer, deit transformer, timesformer, swin transformer and video swin transformer. And we ensemble these models to achieve better results with carefully selected coefficients. Also, we adopt self-training method. Specifically, at first we use the models trained on the downstream datasets to relabel the pre-training videos. Second, we filter the videos which have lower classification confidence than 0.95. Then we resample these videos according to the downstream dataset distribution. Finally, we use a fixed trained teacher and a student network to train the remained videos with downtream videos together. For each iteration, we sample each of them with a batch size ratio 2:1, and optimize KL divergence and cross entropy loss for student network. We observe that some videos have textual information, so we use OCR results to match them with ground truth labels.

[*] Both authors contributed equally to this research.