Unsupervised Multi-View Gaze Representation Learning

John Gideon[1](john.gideon@tri.global), Shan Su[2], Simon Stent[1]
Toyota Research Institute[1], University of Pennsylvania[2]

GAZE2022 4th International Workshop on Gaze Estimation and Prediction in the Wild

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation

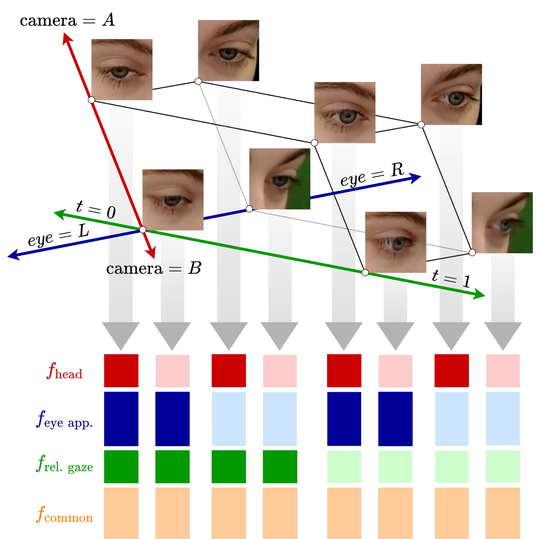
We want to learn gaze estimators with **minimal ground truth supervision** to reduce time and cost. *Sun et al.* introduced the **Cross-Encoder** during ICCV'21, which uses priors about the eyes to **disentangle gaze and appearance**. We expand on their work in the following ways:

Our Contributions

1. Novel feature space that **disentangles head position** from gaze relative to the head
2. A new model, which is more **flexible**, **efficient**, and **performant** than the Cross-Encoder
3. **Interpretability** without any annotation, due to the model confidence output

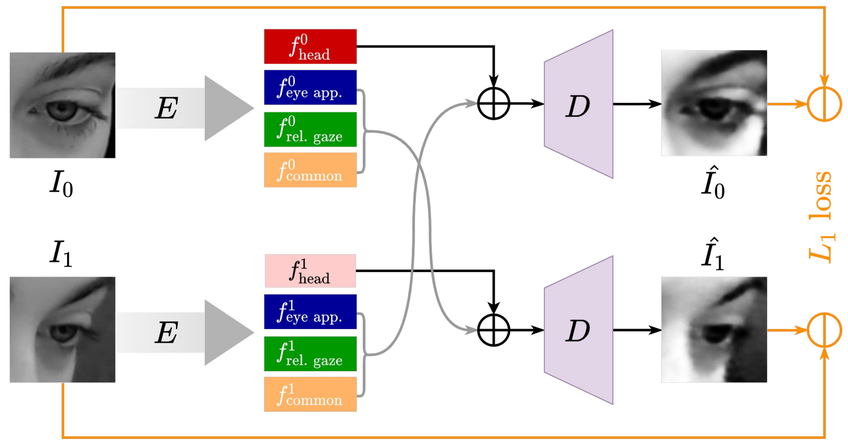
Priors and Sampling Strategy

- **Multi-view:** camera-relative head pose varies depending on the camera position
- **Left-right:** Left eyes share one appearance feature, right eyes another
- **Head-eye dynamics:** Over short intervals of time, the relative gaze (eye motion) changes more than head motion
- **Common factors:** Features related to the subject or overall lighting are consistent over all views



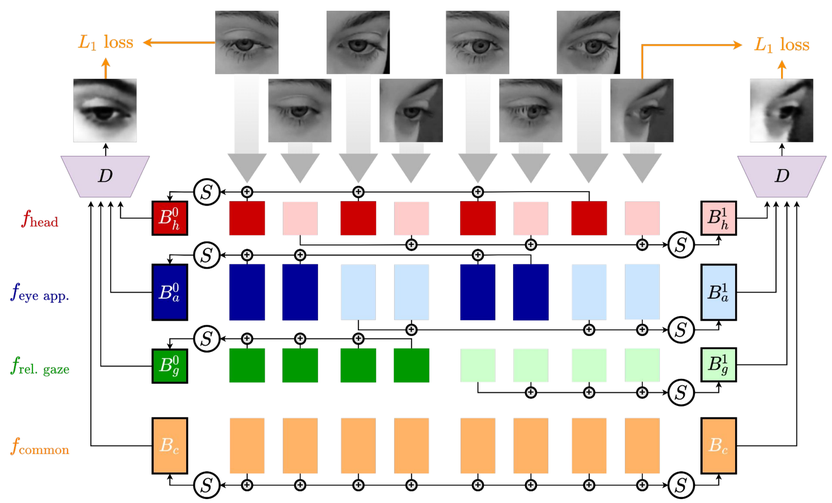
Cross-Encoder with New Features

- Form **three types of pairs** across three input dimensions: (1) camera view, (2) left-right eyes (3) time instances
- Camera view pair is shown below (constant head rotation)
- Since all other feature representations are held constant, the model is able to swap them while preserving the output



Modeling with Basis Features

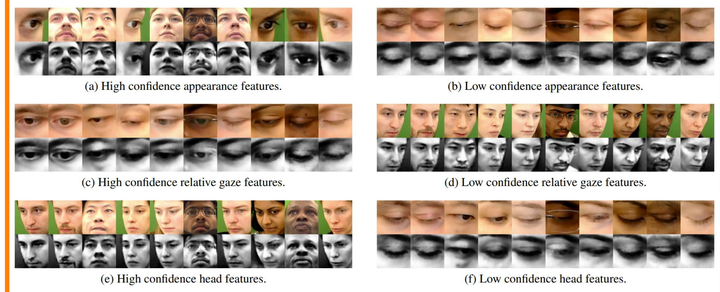
Factors consistent within a subset of views - **basis features**. Basis features can be computed by **summarizing** the views. **Permutations** of basis features reconstruct the input images.



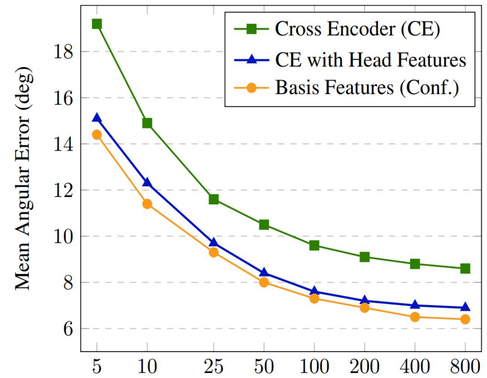
- **Flexible** to missing/extra data during train and test
- **Efficient** - takes half the time to train versus Cross-Encoder

Results

We use weighted accuracy as our summary function to allow for confident views to contribute more. This also provides some interpretability:



Our method yields consistent 2-5° angular improvement vs. the cross-encoder for few-shot gaze estimation on the EVE dataset



Summary

Our structure allows us to disentangle more factors without added annotation cost

The model is flexible, efficient, interpretable, and gives better performance

Code available!
https://github.com/ToyotaResearchInstitute/UnsupervisedGaze