TOYOTA RESEARCH INSTITUTE

Penn University of Pennsylvania

GAZE2022

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

Unsupervised Multi-View Gaze Representation Learning

John Gideon, Shan Su, Simon Stent

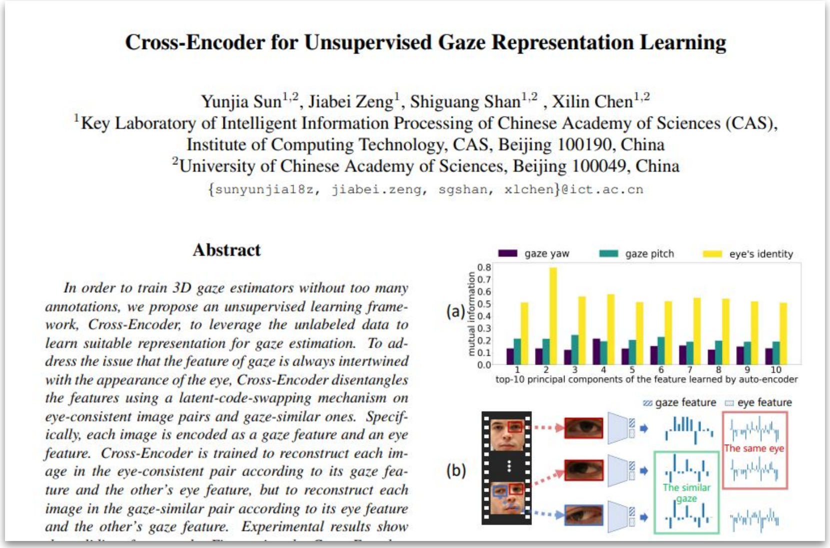
June 2022

Unsupervised gaze representation learning

- Increasing interest to learn gaze estimators with less annotation
- Recently proposed Cross-Encoder uses two pairs of eye images to disentangle gaze and appearance:

Temporal pair
Different gaze
Same appearance

Left-right pair
Same gaze
Different appearance



Cross-Encoder for Unsupervised Gaze Representation Learning

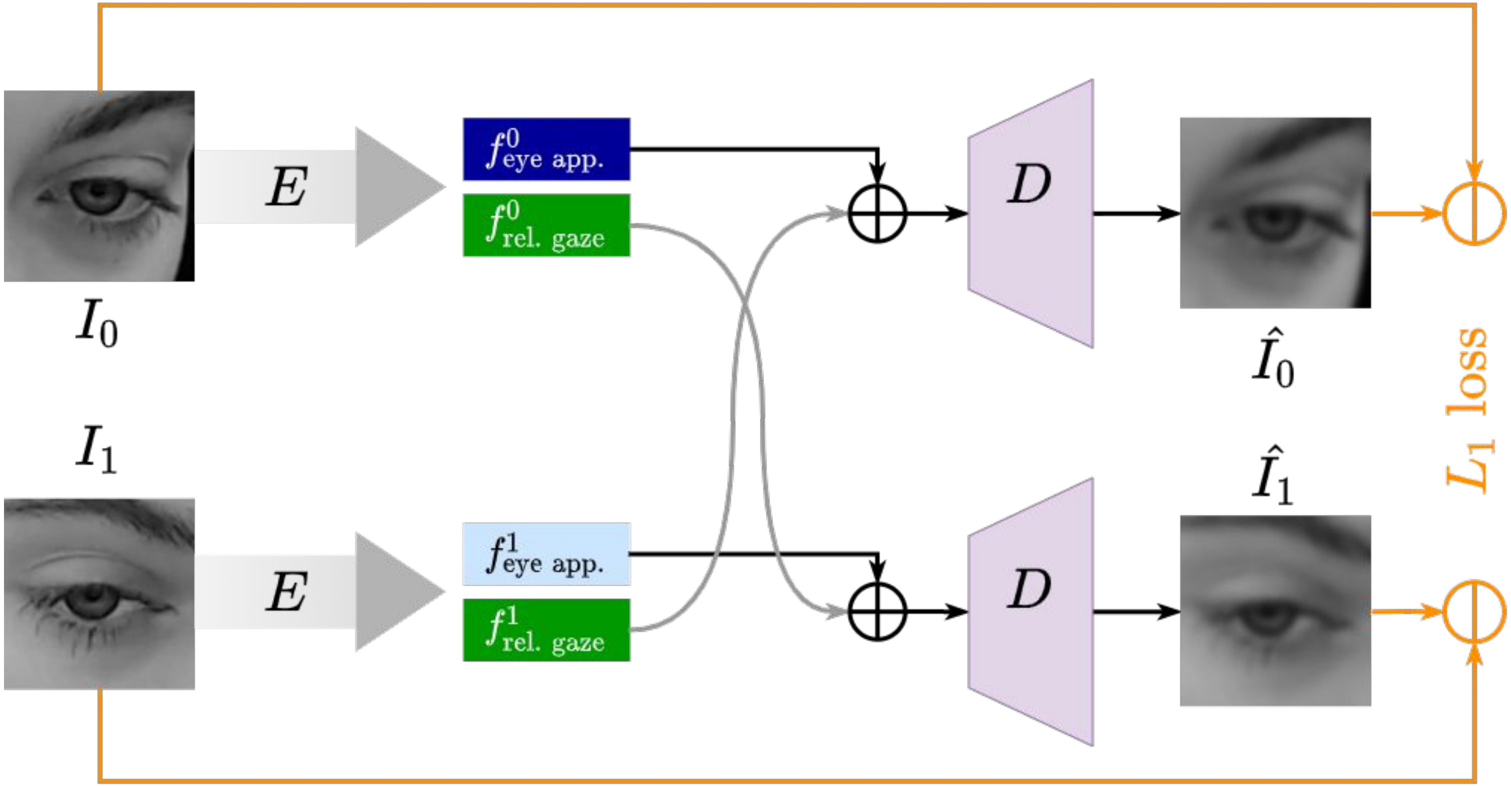
Yunjia Sun[1,2], Jiabei Zeng[1], Shiguang Shan[1,2], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
{sunyunjia18z, jiabei.zeng, sgshan, xlchen}@ict.ac.cn

Abstract

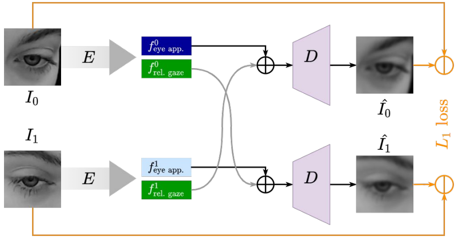
In order to train 3D gaze estimators without too many annotations, we propose an unsupervised learning framework, Cross-Encoder, to leverage the unlabeled data to learn suitable representation for gaze estimation. To address the issue that the feature of gaze is always intertwined with the appearance of the eye, Cross-Encoder disentangles the features using a latent-code-swapping mechanism on eye-consistent image pairs and gaze-similar ones. Specifically, each image is encoded as a gaze feature and an eye feature. Cross-Encoder is trained to reconstruct each image in the eye-consistent pair according to its gaze feature and the other's eye feature, but to reconstruct each image in the gaze-similar pair according to its eye feature and the other's gaze feature. Experimental results show

Sun et al., ICCV 2021

Cross-Encoder Model

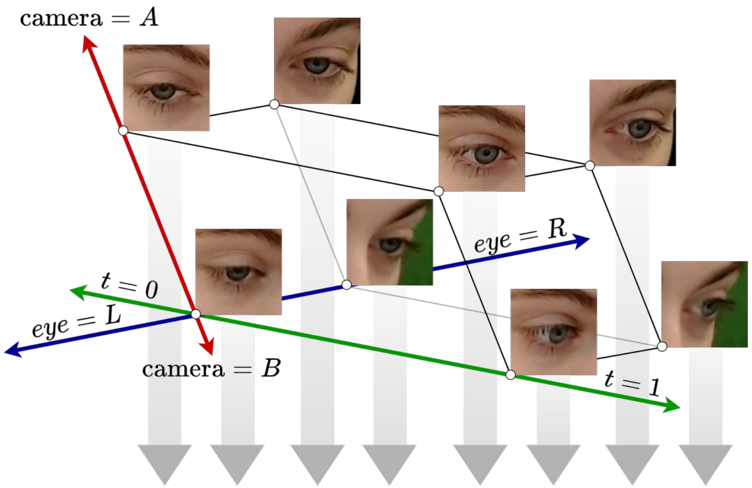


Building on the Cross-Encoder

	Feature Structure	Model Structure	Confidence
Cross-Encoder	Temporal pair Different gaze Same Appearance Left-right pair Same gaze Different Appearance 		
Our Method			

Four weak priors around gaze

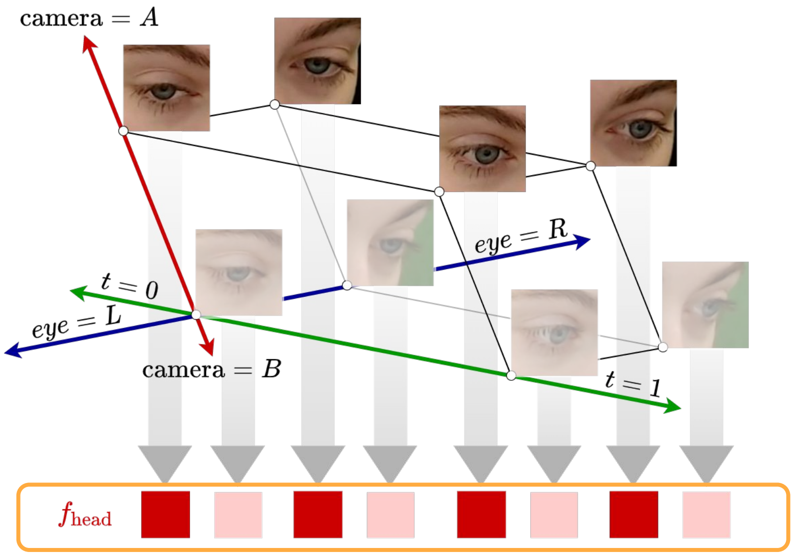
Matrix of eye images across time, camera view, and left/right eye



Four weak priors around gaze

- Multi-view
  - (camera-relative) head pose varies depending on the camera position

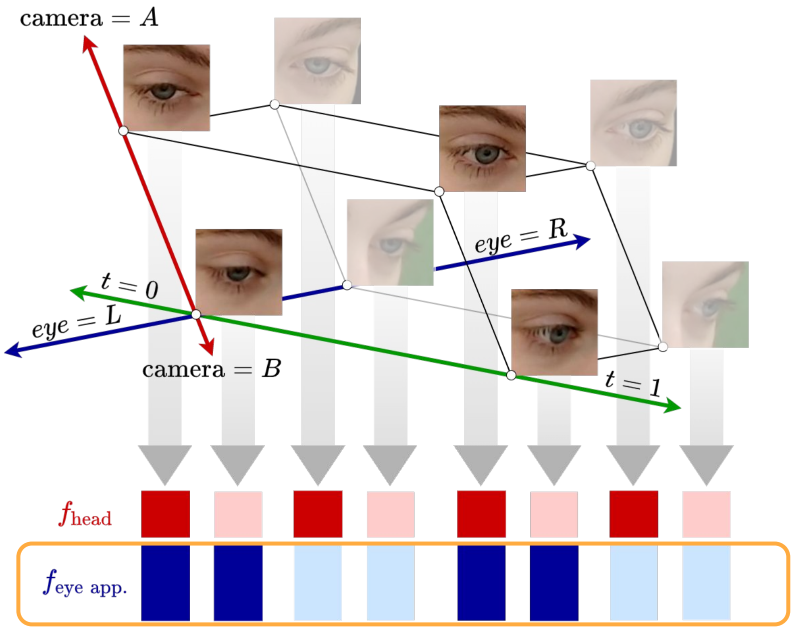
Matrix of eye images across time, camera view, and left/right eye



Four weak priors around gaze

- Multi-view
  - (camera-relative) head pose varies depending on the camera position
- Left-right
  - Left eyes share one appearance feature, right eyes another

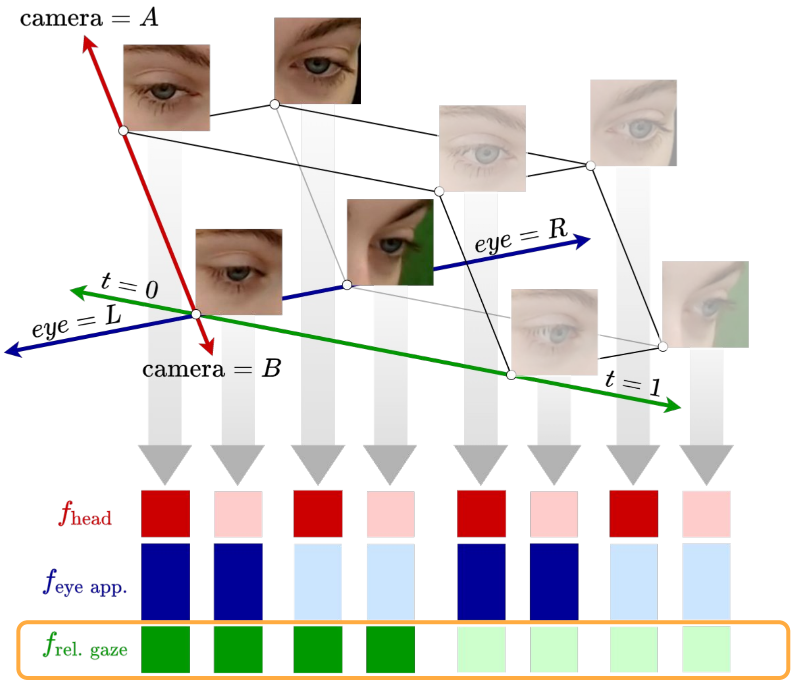
Matrix of eye images across time, camera view, and left/right eye



Four weak priors around gaze

- 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

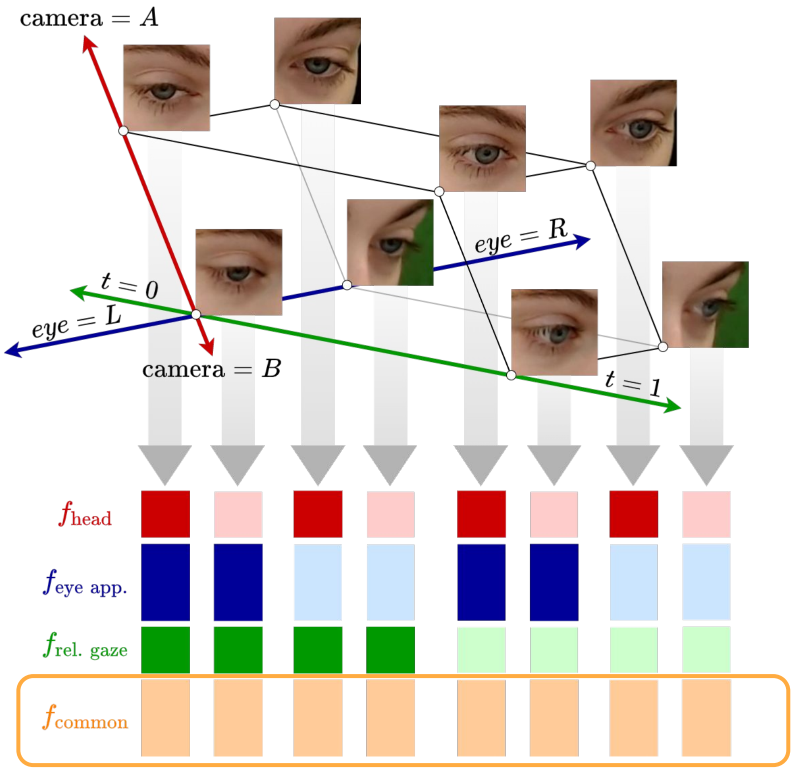
Matrix of eye images across time, camera view, and left/right eye



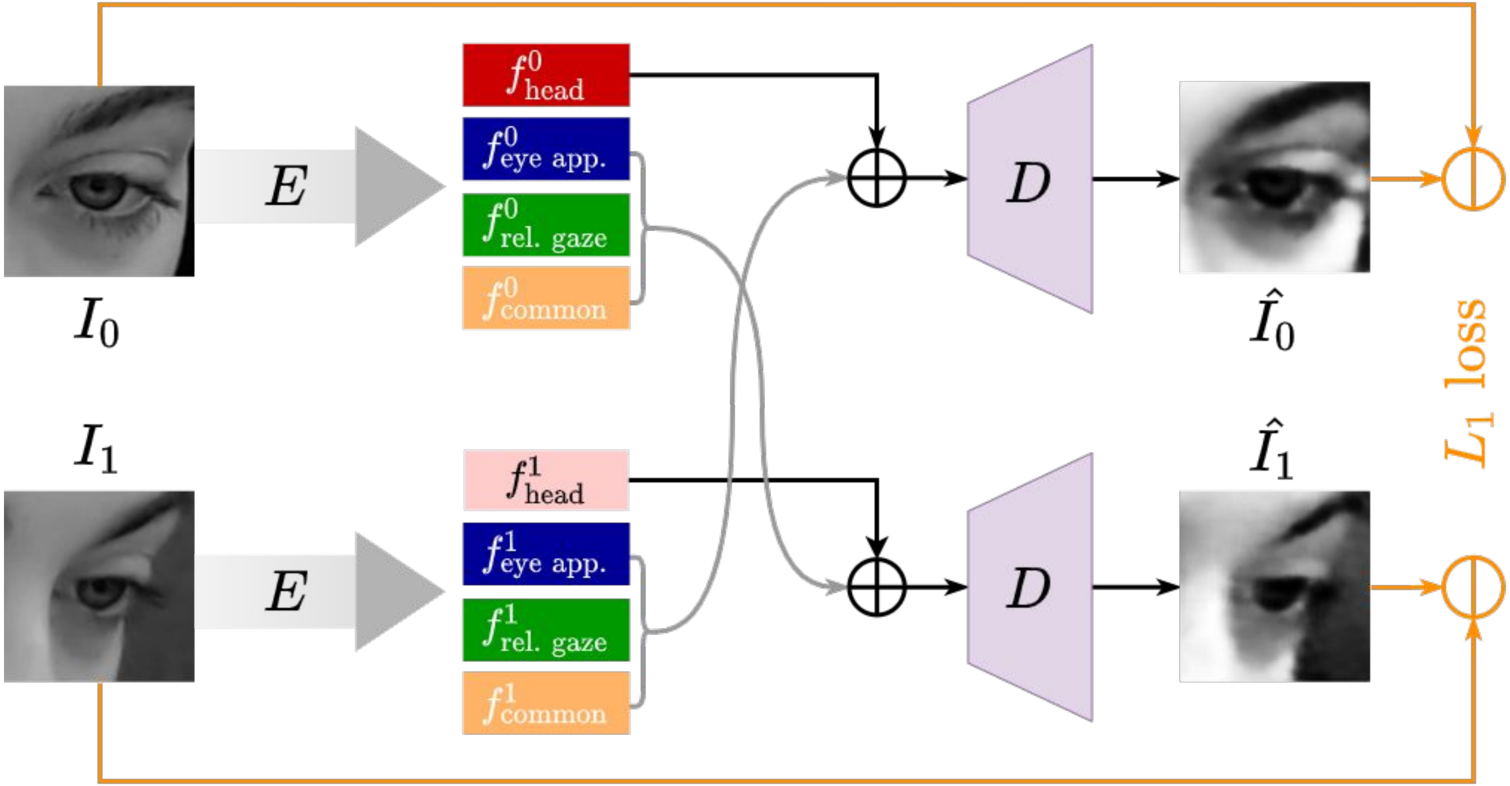
Four weak priors around gaze

- 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

Matrix of eye images across time, camera view, and left/right eye

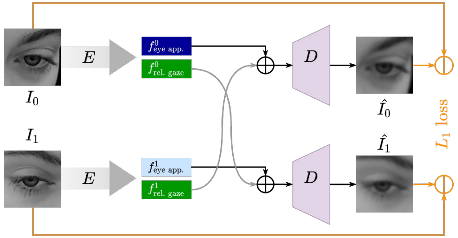
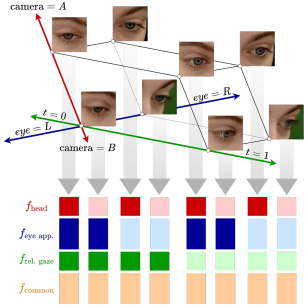


Cross-Encoder with new features



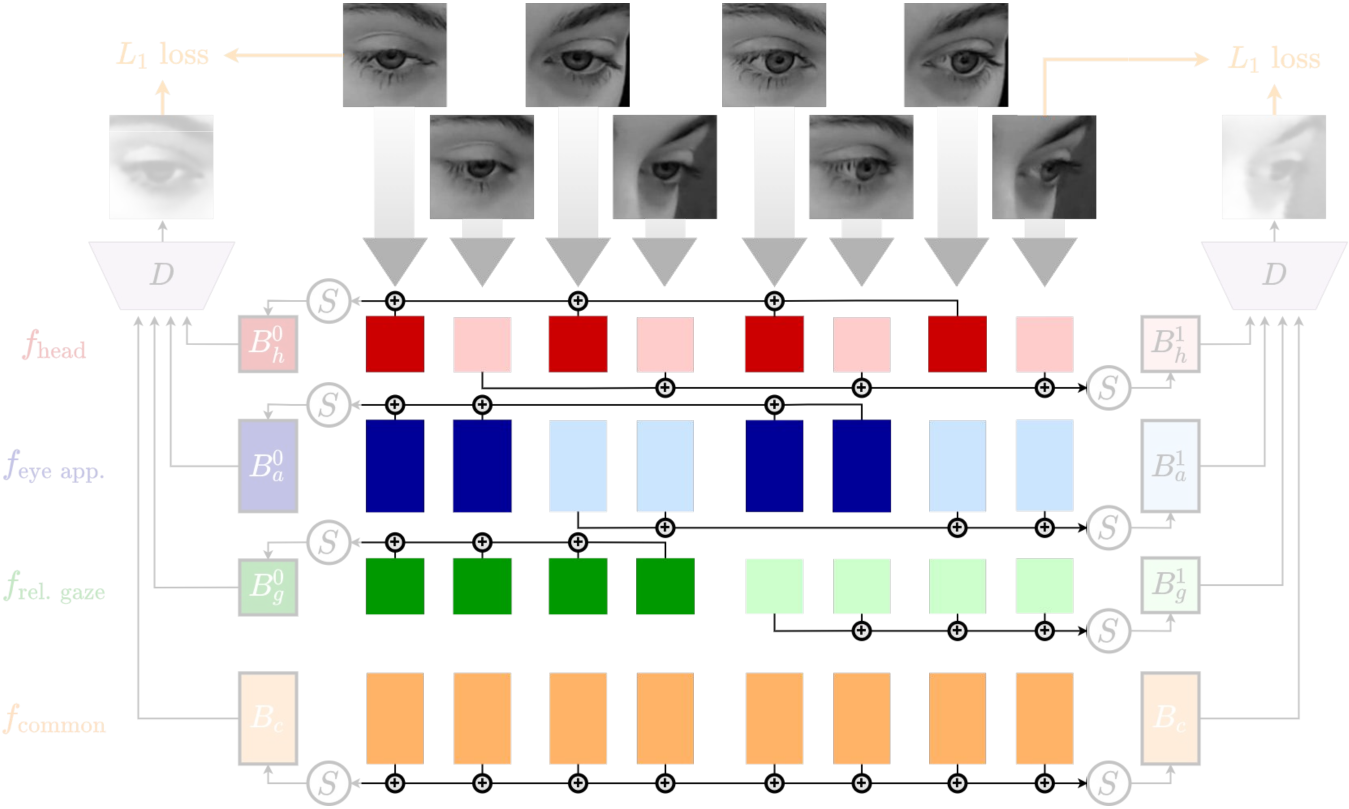
Building on the Cross-Encoder

<table>
<tr><th></th><th>Feature Structure</th><th>Model Structure</th><th>Confidence</th></tr>
<tr><td>Cross-Encoder</td><td>Temporal pair
Different gaze
Same Appearance

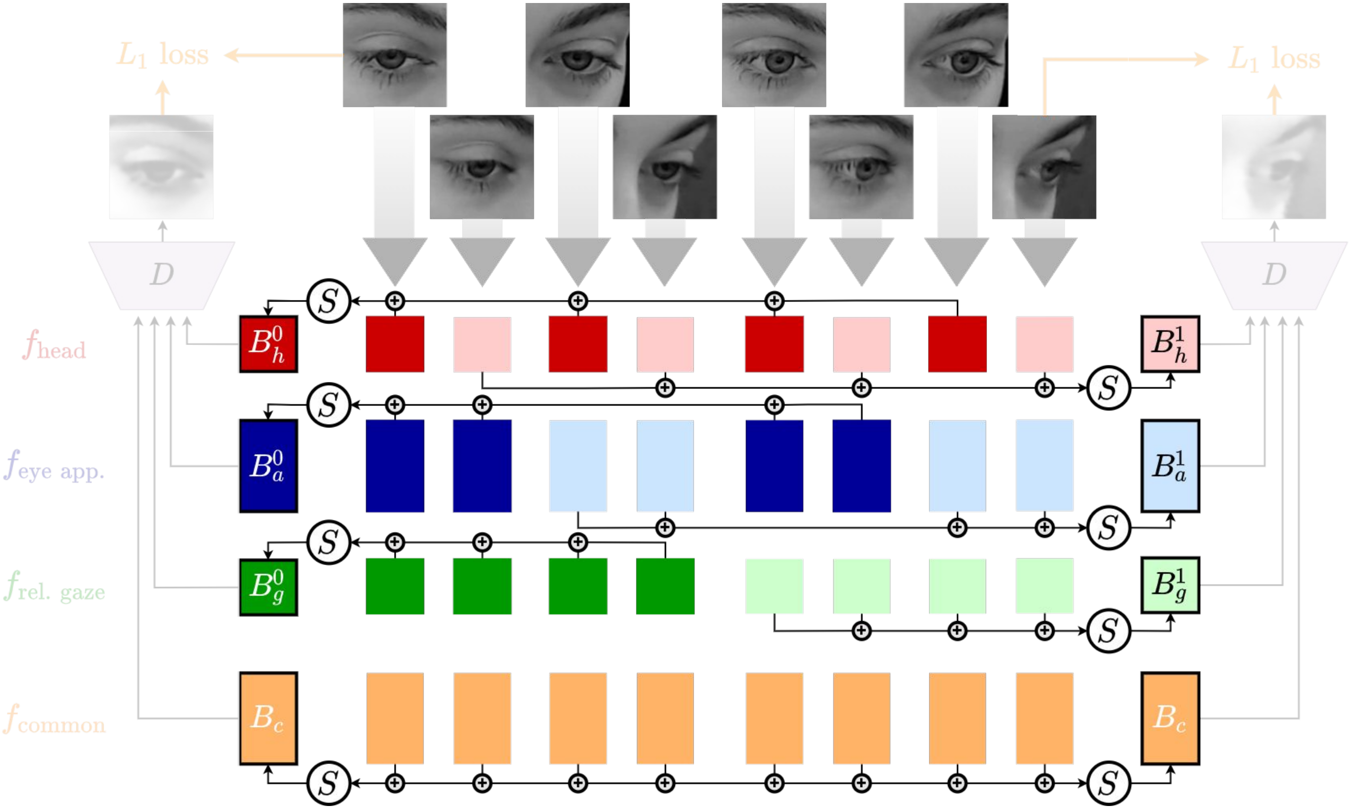
Left-right pair
Same gaze
Different Appearance
</td><td></td><td></td></tr>
<tr><td>Our Method</td><td></td><td></td><td></td></tr>
</table>

Disentangles head rotation from relative gaze

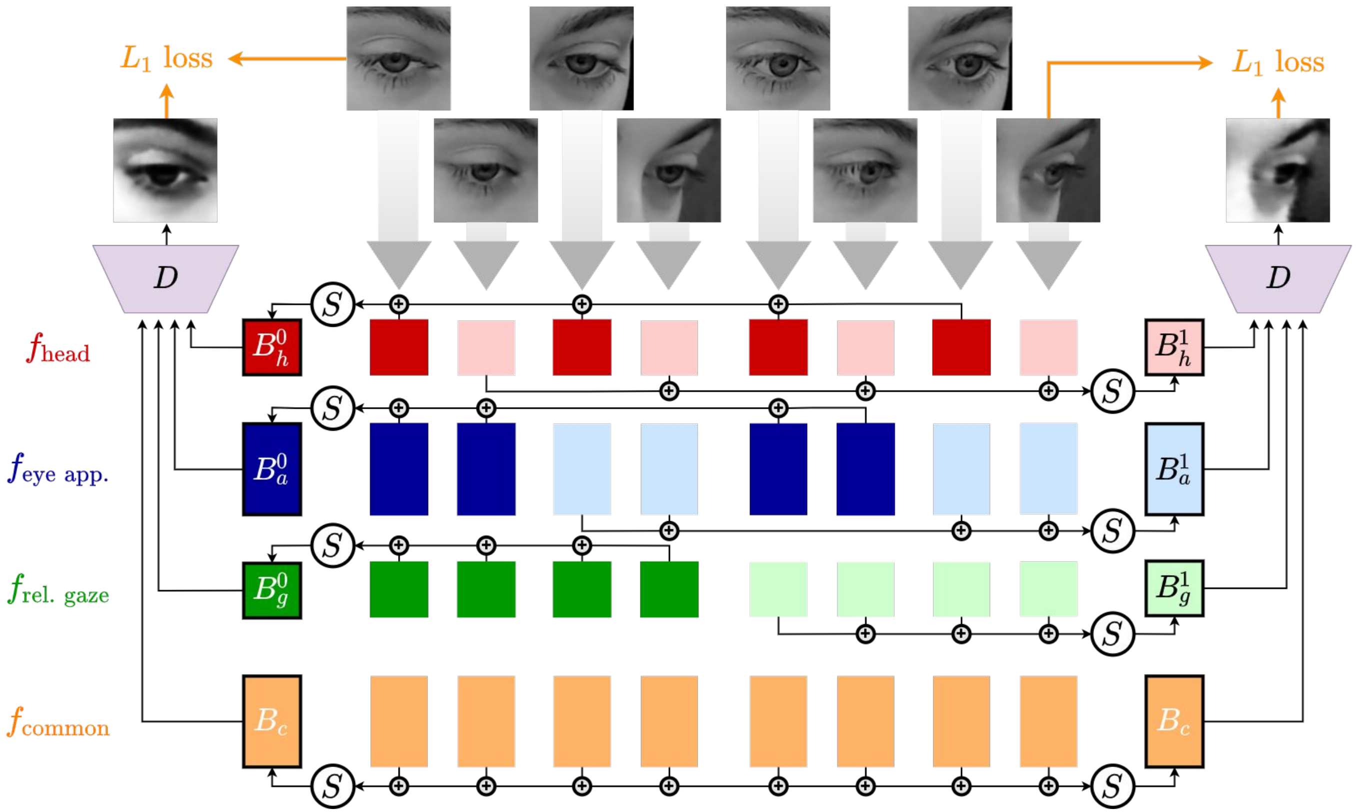
New Model - Basis Loss



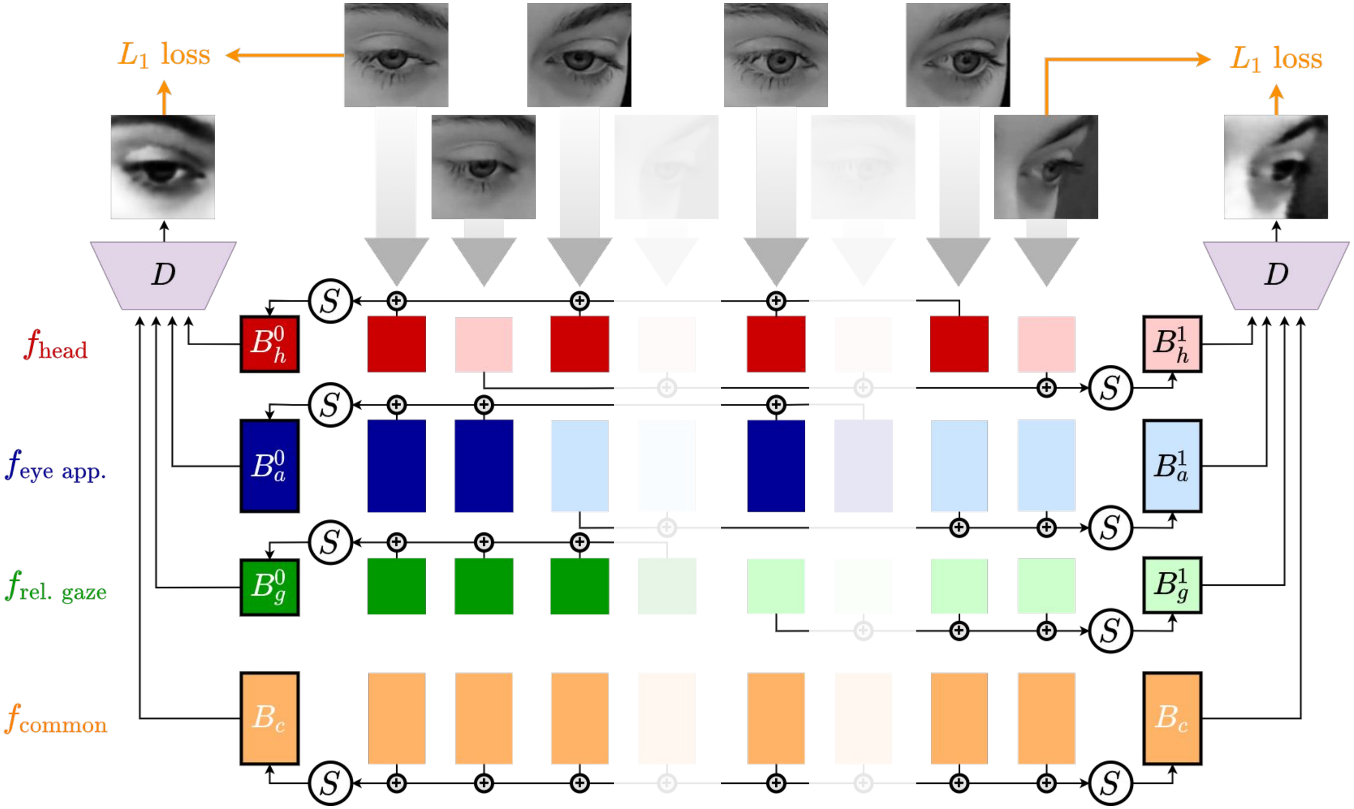
New Model - Basis Loss



New Model - Basis Loss



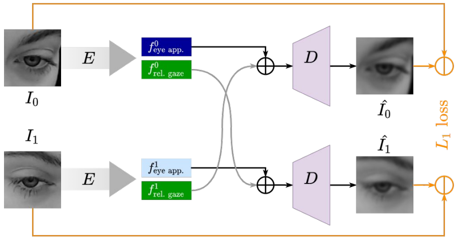
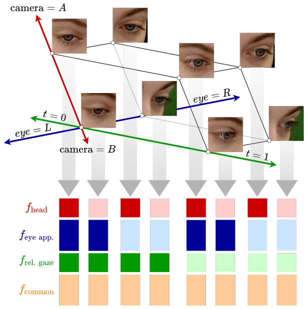
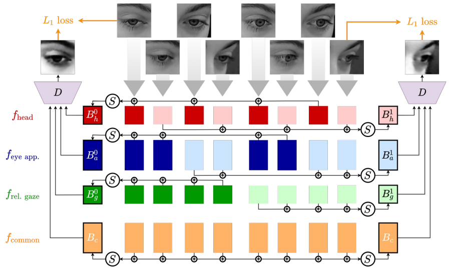
New Model - Basis Loss



<u>Benefits</u>

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

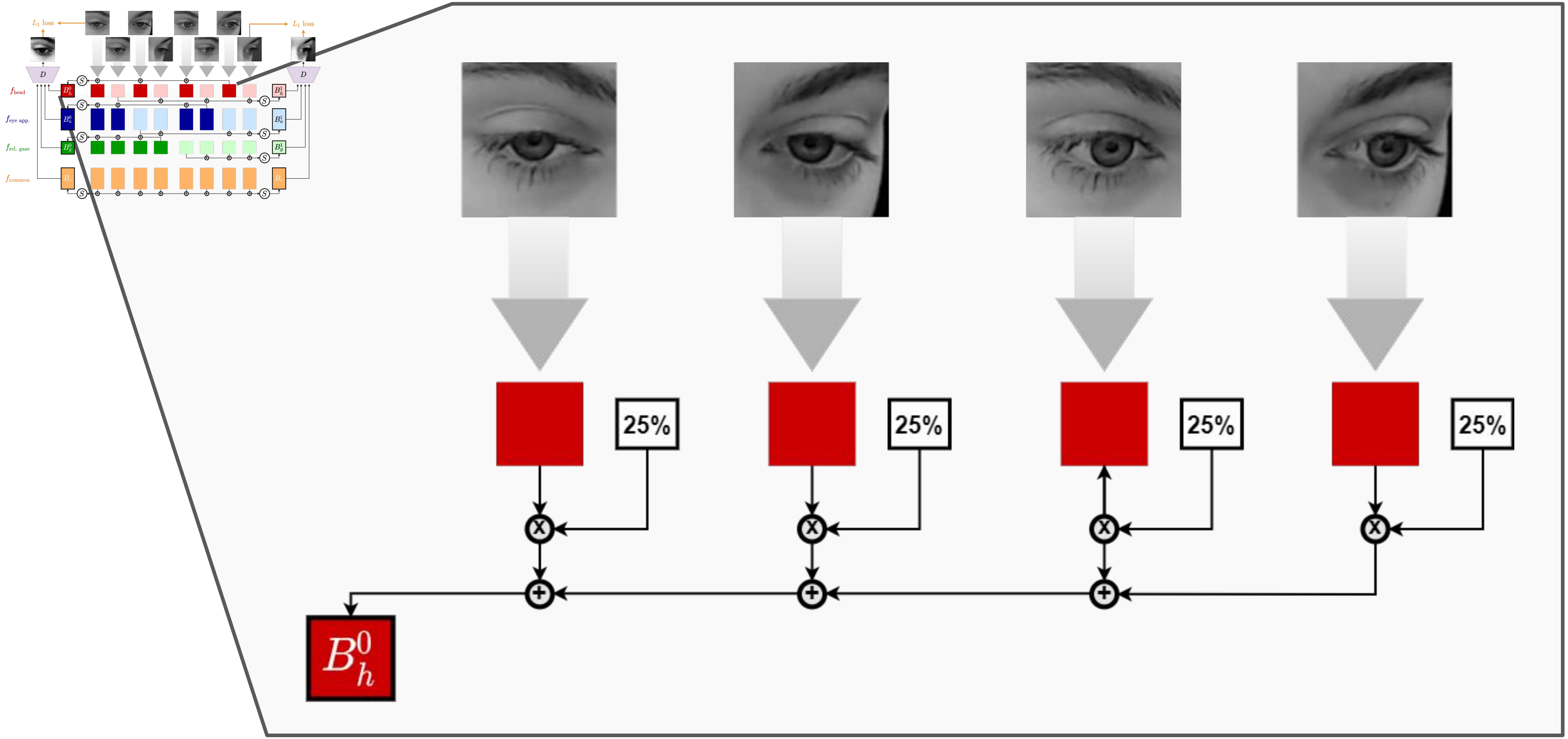
Summary

	Feature Structure	Model Structure	Confidence
Cross-Encoder	Temporal pair Different gaze Same Appearance Left-right pair Same gaze Different Appearance 		
Our Method			

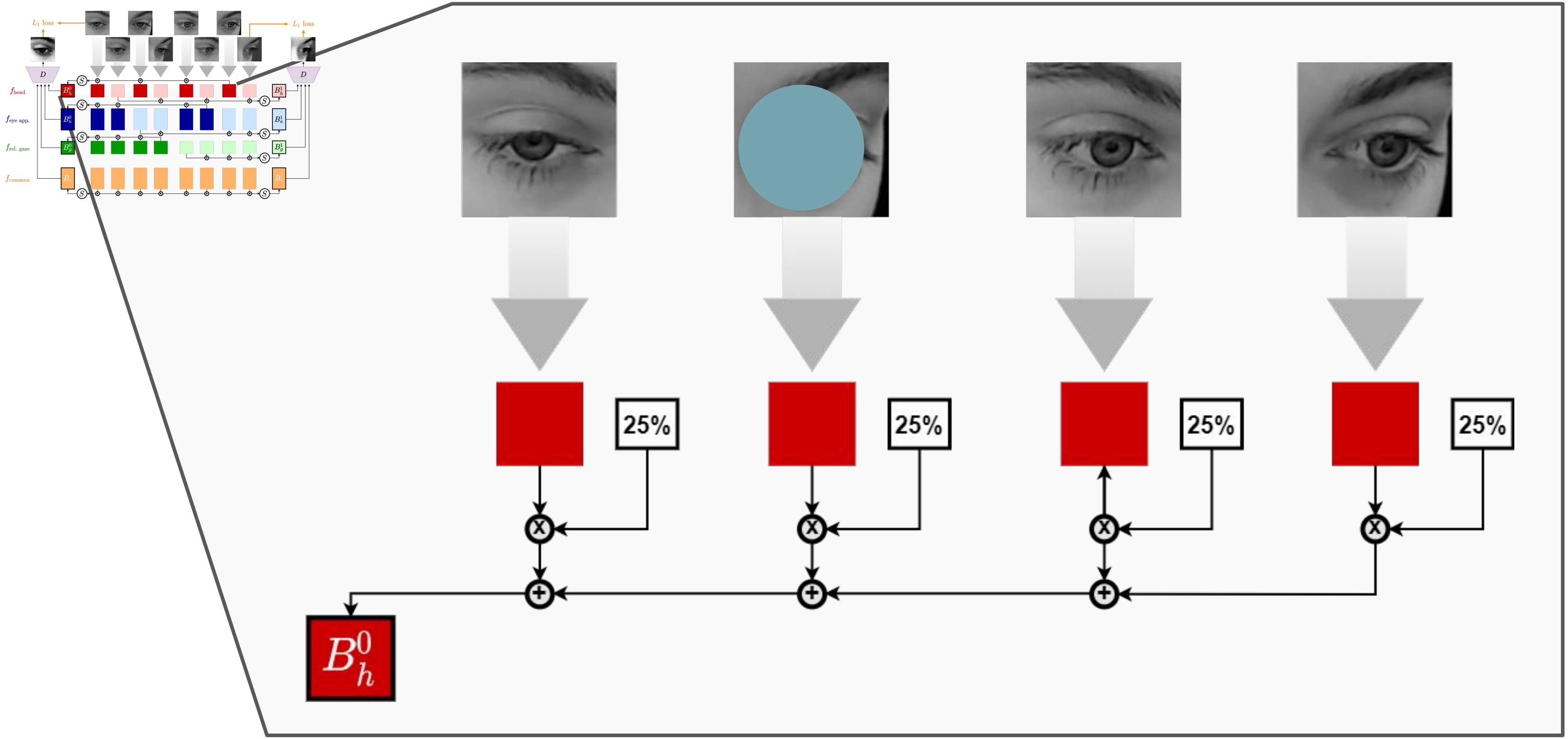
Disentangles head rotation from relative gaze

Flexible and efficient

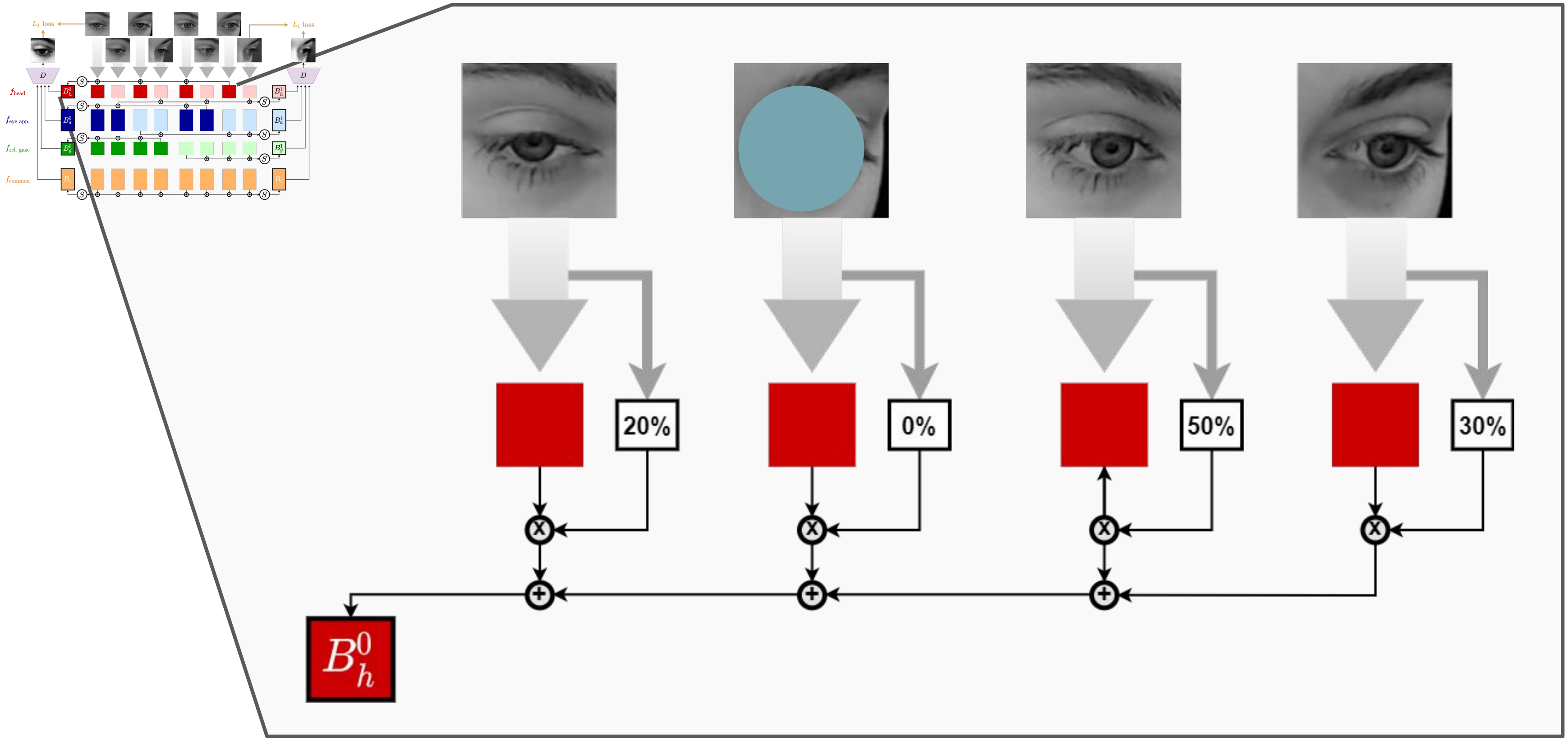
Mean summary function



Mean summary function



Weighting by confidence



Confidence Results

(a) High confidence appearance features.

(b) Low confidence appearance features.

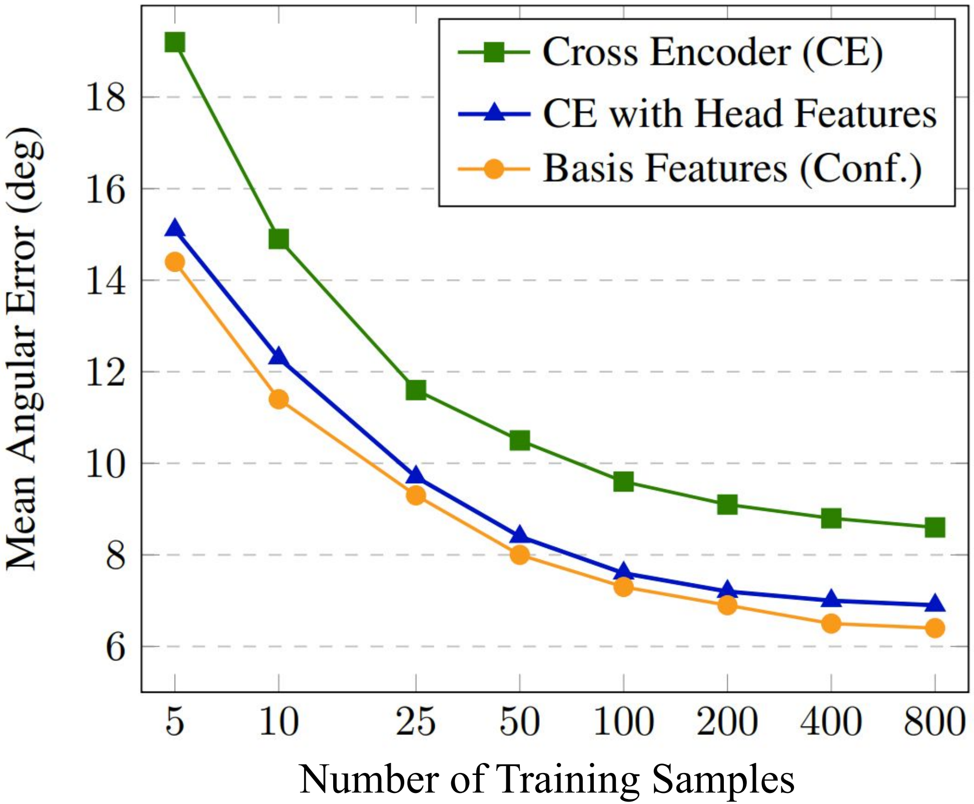
(c) High confidence relative gaze features.

(d) Low confidence relative gaze features.

(e) High confidence head features.

(f) Low confidence head features.

Results



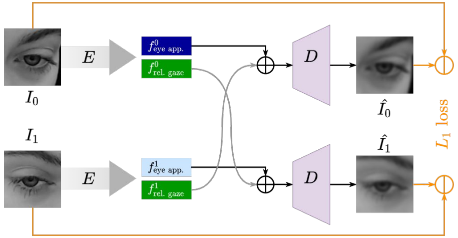
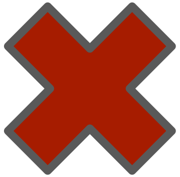
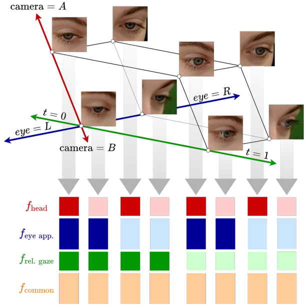
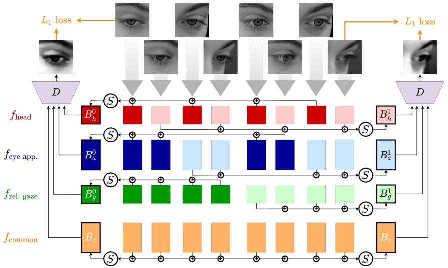
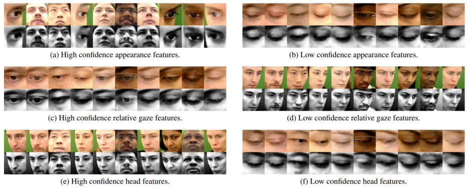
Our method yields consistent 2-5º angular improvement vs. the cross-encoder for few-shot gaze estimation on the EVE dataset

	Without Common	With Common
Mean Baseline	22.7	22.7
Cross Encoder (CE)	**9.6 (0.5)**	12.3 (1.0)
CE with Head Feature	**7.6 (0.3)**	7.8 (0.3)
Basis Loss (mean)	7.9 (0.5)	**7.5 (0.4)**
Basis Loss (confidence)	7.6 (0.5)	**7.3 (0.4)**

For more results, please see our paper

Summary

<table>
<tr><th></th><th>Feature Structure</th><th>Model Structure</th><th>Confidence</th></tr>
<tr><td>Cross-Encoder</td><td>Temporal pair
Different gaze
Same Appearance

Left-right pair
Same gaze
Different Appearance
</td><td>
</td><td></td></tr>
<tr><td>Our Method</td><td>
</td><td>
</td><td>
</td></tr>
</table>

Disentangles head rotation from relative gaze

Flexible, efficient, and performant

Interpretable, even without annotation

Thank You

Code available!

https://github.com/ToyotaResearchInstitute/UnsupervisedGaze