


Problem

Task: to estimate a vector denoting

where the person is looking.



$$g_1 = (x_1, y_1, z_1) = (\theta_1, \varphi_1)$$

Challenges

➤ **Complicated** data collecting procedure → few labeled data.

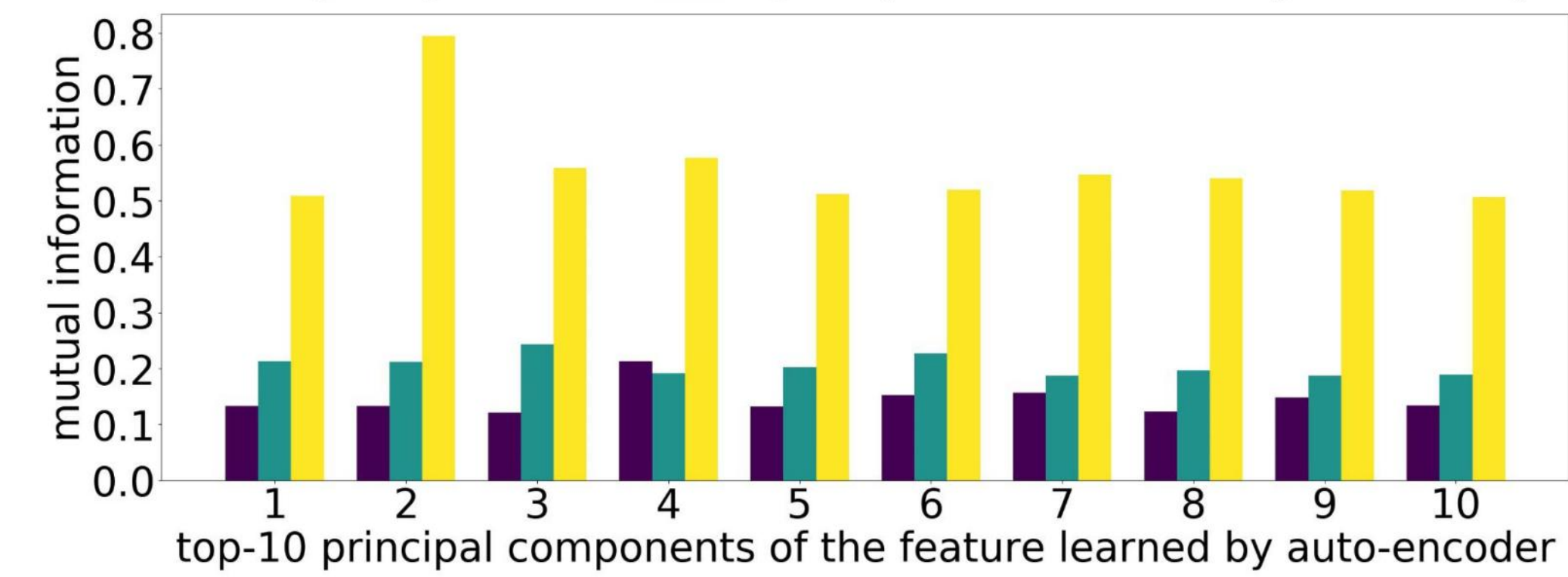
- **unsupervised learning**

➤ General unsupervised feature

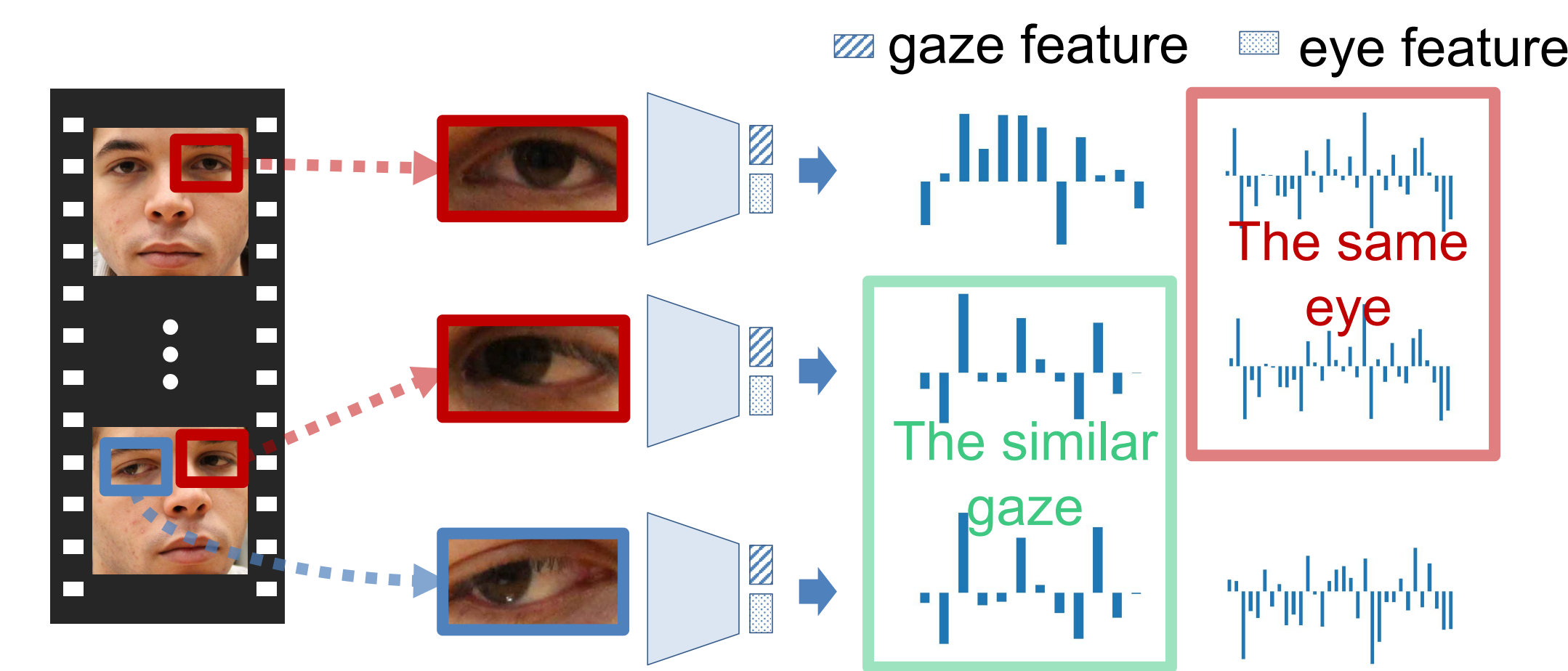
- Eye-identity: **Major** information

Gaze: **Minor** information

gaze yaw gaze pitch eye's identity



How to disentangle gaze and eye identity?

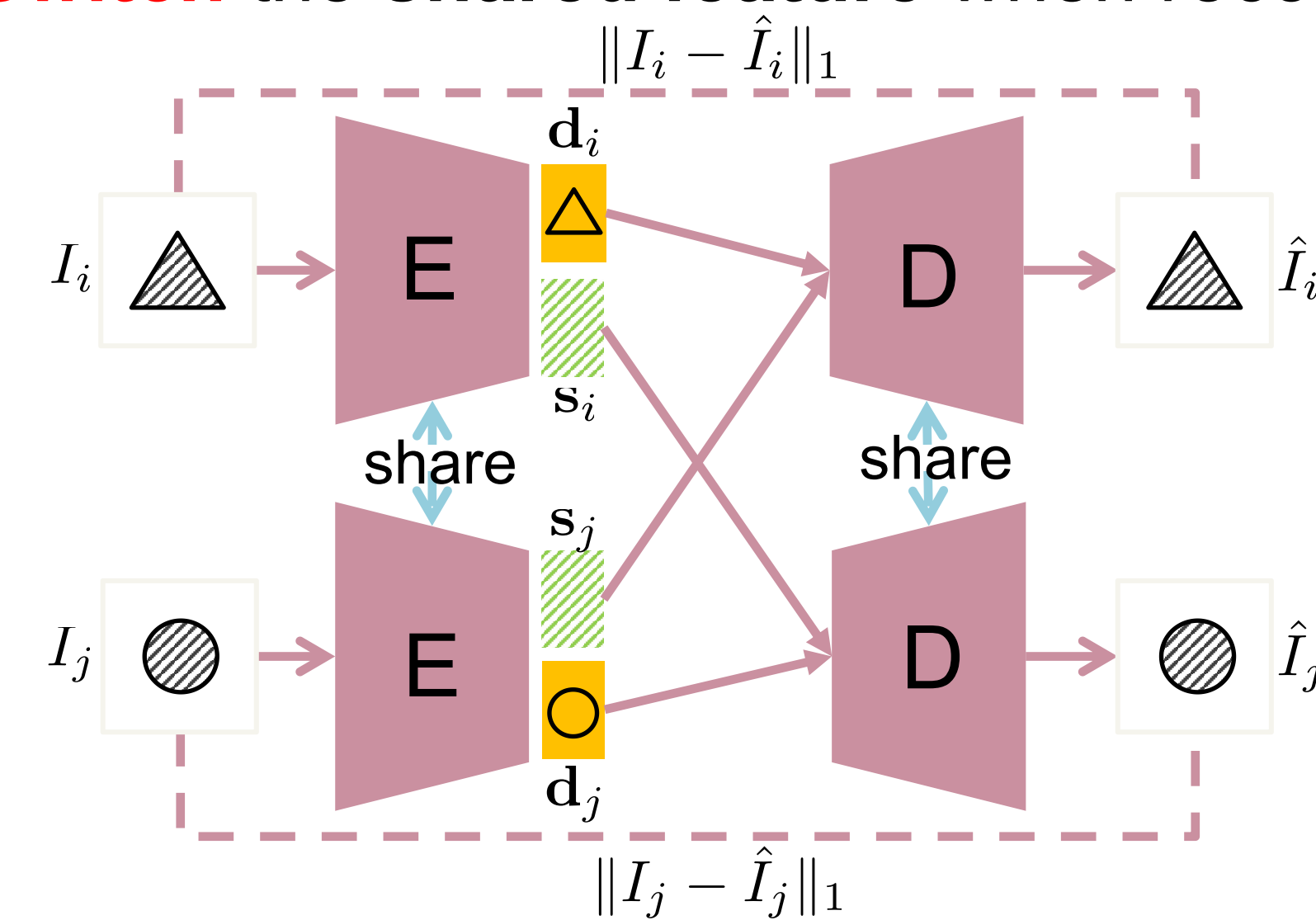


Code: <https://github.com/sunyunjia96/Cross-Encoder>

Cross-Encoder

An extension for auto-encoder.

- Take **a pair of images** as input.
- Each input is encoded in two features.
 - The **shared feature** and the **specific features**.
- **Switch** the shared feature when reconstruction.



Cross-Encoder for unsupervised gaze representation learning

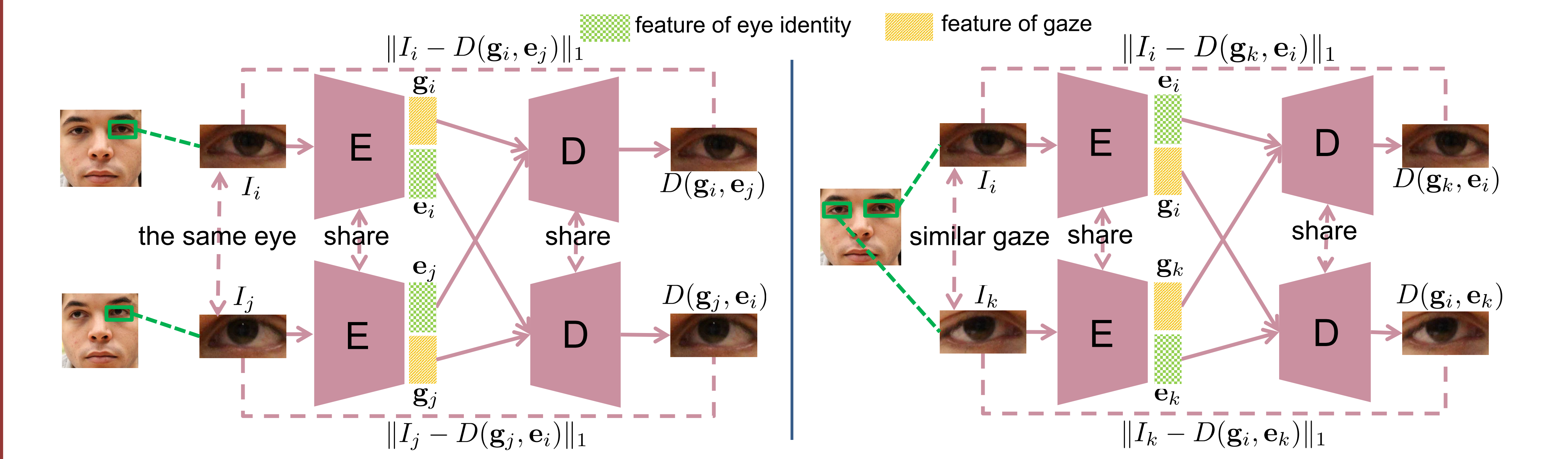
An eye image is encoded in the **gaze feature** and the **eye feature**.

Two kinds of image pairs to avoid the degenerative solution.

Eye-consistent pair: paired images of the same eye.

Gaze-similar pair: paired images of the left and the right eyes in an image.

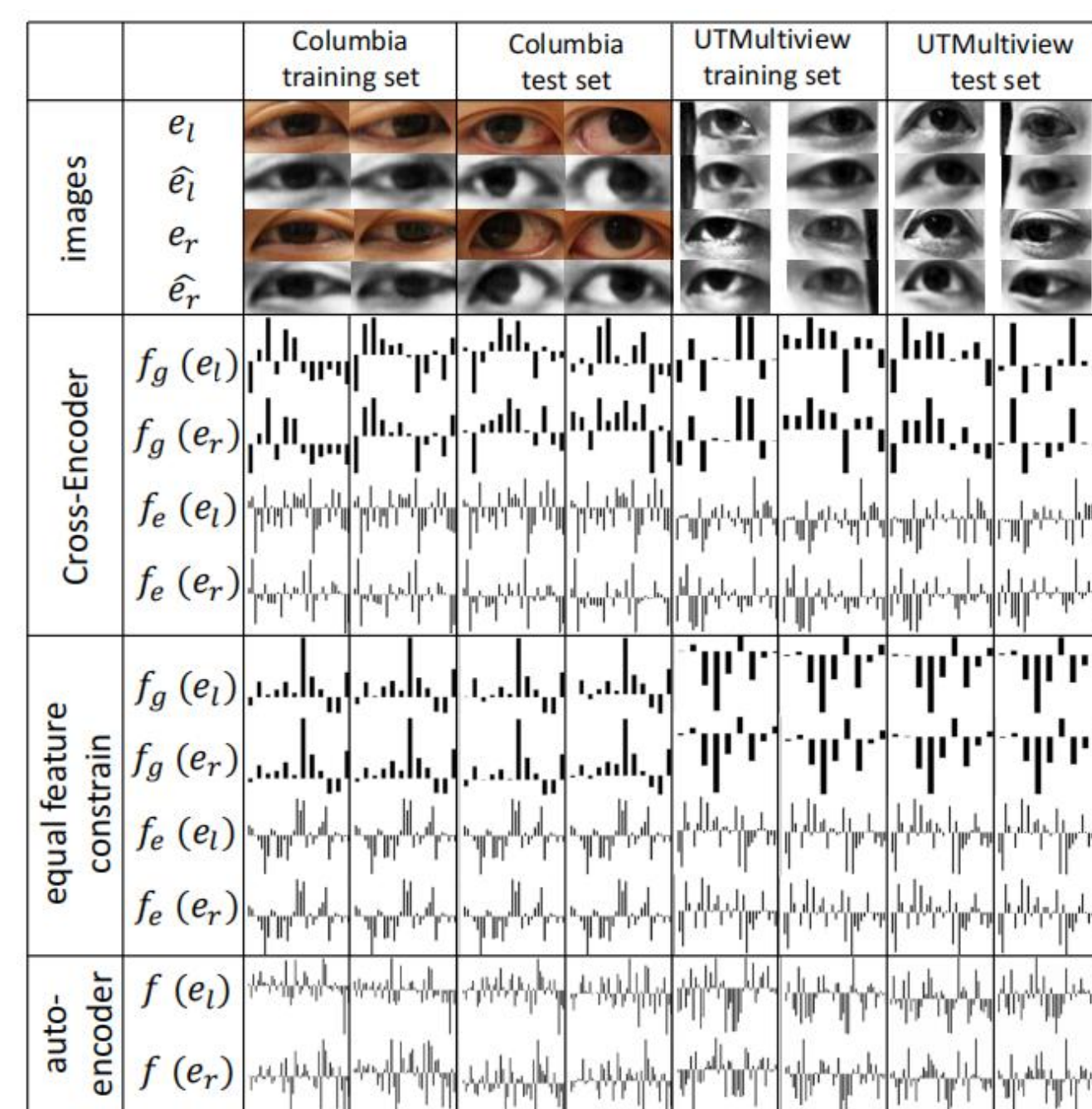
Same: **eye identity** Different: **gaze** ○ Switch: feature of **eye identity e** Same: **gaze** Different: **eye identity** ○ Switch: feature of **gaze g**



Experiments

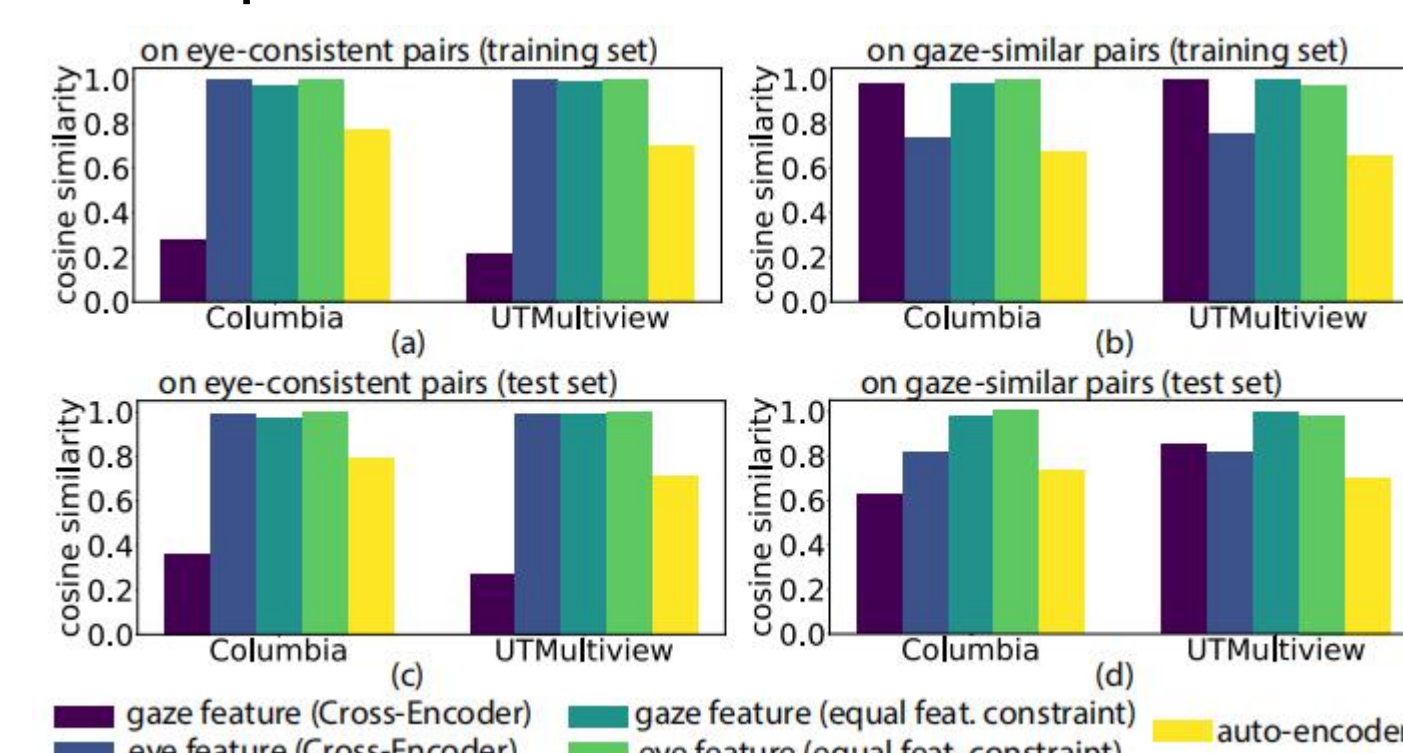
Disentanglement of features

➤ Examples of the learned representations.

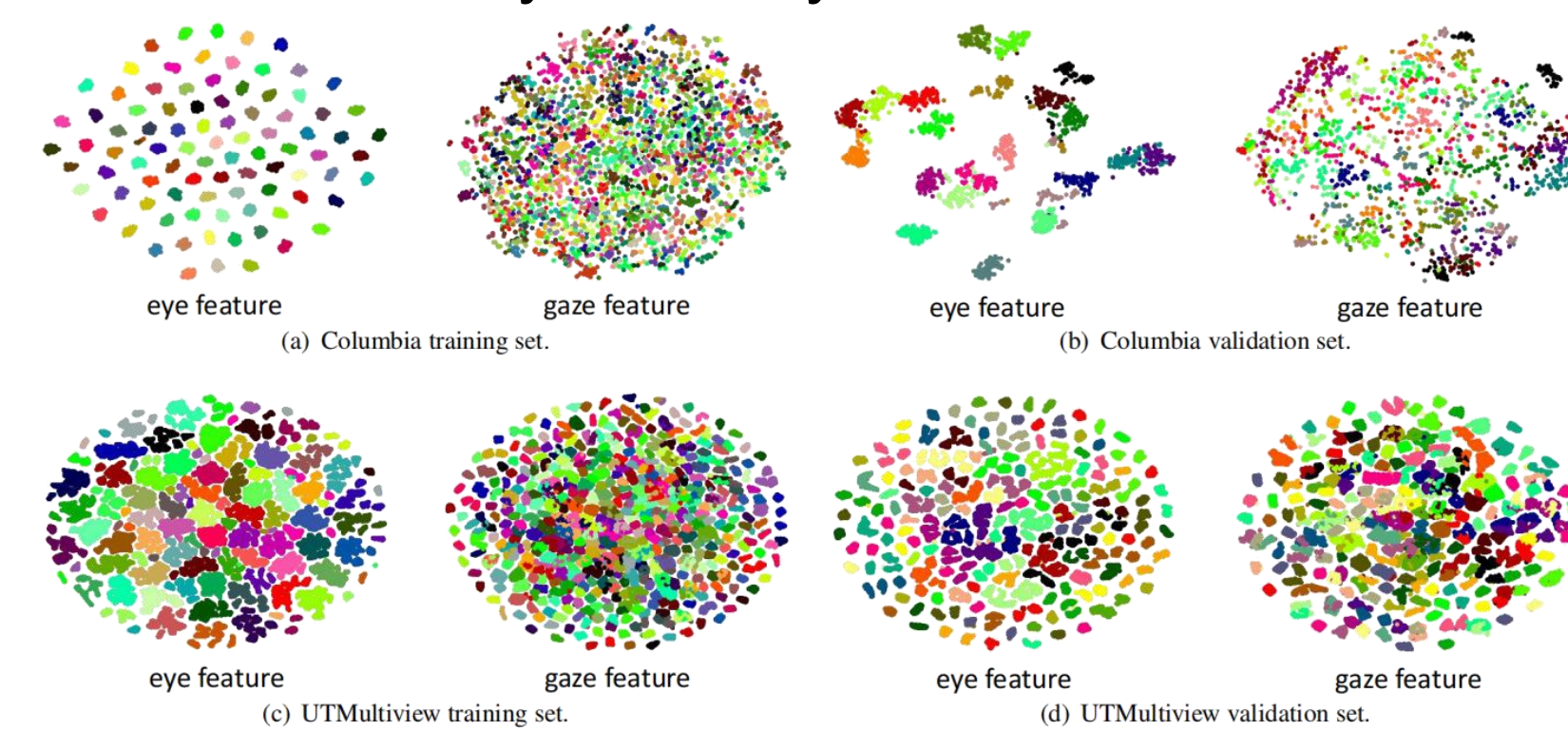


e_l and e_r are the left and the right eye. \hat{e}_l and \hat{e}_r are their reconstructed images. f_g is the gaze feature, and f_e is the eye feature.

Cosine similarity of the learned representations.



T-SNE results of the learned representations, colored in eye-identity.



Within dataset evaluation

➤ Cross-Encoder outperforms the state-of-the-art consistently.

Table 2. Angular errors of 50/200-shot gaze estimation within datasets.

		AE(EFC)	simCLR	BYOL	ours
50	C	11.1±0.3	8.1±0.04	10.8±0.1	7.0±0.2
	U	14.9±0.5	14.4±0.5	15.1±0.5	8.8±0.4
	M	9.8±0.2	10.7±0.4	11.9±0.4	8.5±0.2
200	C	7.8±0.1	6.3±0.02	9.4±0.03	6.2±0.1
	U	11.9±0.3	11.0±0.3	14.1±0.2	7.3±0.2
	M	8.8±0.1	9.2±0.3	10.4±0.4	7.3±0.1

Table 1. Angular errors of 100-shot gaze estimation within datasets.

methods	w/ head pose			w/o head pose		
	Columbia	UTMultiview	MPIIGaze	Columbia	UTMultiview	MPIIGaze
ImageNet-Pretrained ResNet18	12.1±0.1	20.2±0.5	10.6±0.2	11.9±0.2	24.9±0.5	10.6±0.2
auto-encoder	10.5±0.2	18.0±0.5	9.5±0.2	10.6±0.3	18.5±0.5	9.5±0.1
auto-encoder (EFC)	9.2±0.3	13.5±0.3	9.2±0.2	9.4±0.3	22.1±0.5	8.9±0.1
SimCLR ICML '20	7.2±0.1	12.1±0.2	10.0±0.3	8.2±0.03	21.3±0.7	9.8±0.2
BYOL NIPS '20	9.9±0.1	14.4±0.2	11.1±0.5	10.2±0.03	23.5±0.2	11.0±0.6
Yu et al. CVPR '20	8.95	8.56	-	-	-	-
Cross-Encoder (proposed)						
- eye feature	12.8±0.1	15.5±0.4	9.8±0.1	12.6±0.2	31.9±0.3	9.7±0.1
- gaze feature (no GS pair)	7.6±0.1	10.6±0.3	8.2±0.1	8.5±0.2	17.2±0.6	8.1±0.2
- gaze feature (no residual loss)	6.7±0.1	7.4±0.1	7.2±0.2	7.4±0.1	8.2±0.2	7.2±0.2
- gaze feature ($d_g=9, d_e=32$)	6.7±0.1	7.7±0.3	8.1±0.2	7.6±0.1	8.8±0.2	8.0±0.2
- gaze feature ($d_g=12, d_e=32$)	6.6±0.1	8.0±0.2	7.5±0.1	7.3±0.1	8.9±0.2	7.6±0.2
- gaze feature ($d_g=15, d_e=32$)	6.4±0.1	8.0±0.2	7.5±0.2	7.1±0.1	9.2±0.2	7.3±0.2
- gaze feature ($f_g=12, f_e=16$)	6.7±0.1	7.6±0.2	7.2±0.2	7.4±0.2	8.6±0.2	7.2±0.1
- gaze feature ($f_g=12, f_e=64$)	6.5±0.1	7.8±0.2	7.5±0.2	7.2±0.1	8.9±0.1	7.4±0.1

Cross dataset evaluation

➤ Cross-Encoder outperforms the state-of-the-art consistently.

Table 3. Angular errors of 100-shot gaze estimation cross datasets.

	C	U	M
supervised			
- trained on C	-	10.84	8.35
- trained on U	7.19	-	8.11
- trained on X	5.67	8.79	7.28
unsupervised			
- Yu et al. CVPR '20 (trained on U)	8.82	-	-
Cross-Encoder			
- trained on C	-	9.79	8.32
- trained on U	7.48	-	9.09
- trained on X	7.76	10.30	9.04
- trained on X, T, and F	7.09	9.58	8.20

Fine-tune within datasets

➤ Cross-Encoder is competitive with the state-of-the-art.

Table 4. Angular errors of the state-of-the-art gaze estimation methods and Cross-Encoder as a pretrained model.

	Columbia	UTMultiview
Yu et al. CVPR '20	3.42	5.52
Park et al. ECCV '18	3.59	-
Zhang et al. CVPR '15	-	5.9
Wang et al. CVPR '19	-	5.4
Cross-Encoder(proposed)	3.52	4.81