# Supplementary Material for
# *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

We provide additional experimental results that are not included in the main manuscript due to space limitation. We present ablation study about different components in Section 1, and show additional visualization results in Section 2.

## 1. Discussion of different components

We discussed the effectiveness of the eye-consistent pair, gaze-similar pair, and residual loss by comparing the variants of Cross-Encoder where the components are omitted in turns. When all the components are omitted, the Cross-Encoder is degenerated to the vanilla auto-encoder. Table 1 reports the mean angular errors of 100-shot gaze estimation with different components. Below are our observations.

**Eye-consistent pair:** The eye-consistent pair plays the leading role in Cross-Encoder. Comparing the results in the 1-st row with those in the 2-nd row, the Cross-Encoder using eye-consistent pair surpasses auto-encoder. This is because compared to the auto-encoder, using eye-consistent pair leads the Cross-Encoder to encode gaze information and wipes off unrelated factors in the gaze feature. We also observe that, if only one pair is used, the Cross-Encoder using eye-consistent pair(the 2-nd row) significantly outperforms the one using the gaze-similar pair(the 4-th row). This is because when the eye-consistent pair is used, the Cross-Encoder encodes the shared eye's information into the switchable feature and leaves almost all the gaze information in the fixed feature. However, the gazes of the two images in gaze-similar pair are similar rather than being absolutely equivalent. When only the gaze-similar pair is used, the Cross-Encoder might leave some of the gaze information in the eye feature. When comparing the 4-th row with the 6-th row, or the 5-th row with the 7-th row, we can see that adding the eye-consistent pair significantly reducing the errors.

**Gaze-similar pair:** Although using a single gaze-similar pair leads to poor performance, using the gaze-similar pair together with the eye-consistent pair boosts the performance of Cross-Encoder. Comparing the 4-th and 5-th row with the 1-st row, it can be seen that the results of only using gaze-similar pair is worse than that of the auto-encoder. The reason is that the gazes between images in gaze-similar pairs are not absolutely equivalent so that the Cross-Encoder might encode partial gaze information into the eye feature rather than the gaze feature. Nevertheless, comparing the 2-nd and 3-rd row with the last two rows, we can see that adding gaze-similar pair further improves the performance of the Cross-Encoder. The reason is that, if only eye-consistent pair is used, the Cross-Encoder might degenerate to the solution that it encodes all information (including the eye information) in the gaze feature. Adding gaze-similar pair forces the Cross-Encoder to pull the eye information out from the gaze feature, thus avoiding the degenerated solution.

**The residual loss:** It can be seen that the residual loss helps when only eye-consistent pair exists when we compare the 2-nd row with the 3-rd row. Recall that in the main manuscript, the residual loss helps to prevent the Cross-Encoder to encode the difference between the input pair into the shared features, thus keeping the gaze information in the gaze feature integrated.

When comparing the 4-th row with the 5-row, we observe that the residual loss has conflicted effect on the performance if only the gaze-similar pair is used. The reason is that, in this case, the gaze feature is the shared feature. The residual loss only prevents the eye information to be encoded in the gaze feature, but cannot ensure that the gaze information is completely encoded in the gaze feature.

Comparing the 6-th row and the 7-th row, the effect of the residual loss is ambiguous too. Because when both pairs are used, minimizing the two reconstruction losses is sufficient to prevent the information leaking to the other feature. Also as the distribution of the two datasets differs, the residual loss does not show significant enhancement. We concluded that residual loss is optional and the reconstruction loss is

Table 1. Mean angular errors of 100-shot gaze estimation with different components on Columbia and UTMultiview. On both the two datasets, we used the learned gaze features with the optimal dimension and concatenated with the head pose. The first row without any components denotes the vanilla auto-encoder.

| residual loss | gaze-similar pair | eye-consistent pair | Columbia | UTMultiview |
|---|---|---|---|---|
| × | × | × | 9.35 | 17.78 |
| × | × | ✓ | 7.62 | 11.32 |
| ✓ | × | ✓ | 7.44 | 10.51 |
| × | ✓ | × | 11.68 | 15.00 |
| ✓ | ✓ | × | 11.76 | 14.63 |
| × | ✓ | ✓ | 6.76 | **7.34** |
| ✓ | ✓ | ✓ | **6.51** | 7.71 |

the cutting edge loss.

## 2. Additional visualization results

We provide extra visualization results to demonstrate the quality of the features learned by the Cross-Encoder.

Figure 1 shows other examples of the learned representations. It also leads to the conclusion that the gaze feature and eye feature learned by Cross-Encoder is disentangled, as the gaze feature of the gaze-similar pair stays similar and the eye feature of the gaze-similar pair varies. Also, the eye feature of the eye-consistent pair stays very alike and the gaze feature of the eye-consistent pair is different. The features learned by equal feature constrained auto-encoder stays almost the same all the time.

Figure 2 shows the eye features and gaze features in 2-dimensional space using t-SNE[1]. Each point denotes the feature of an eye's image. The points are colored by their gaze labels. In particular, we transfer the gaze label into normalized 3-dimensional vector $\boldsymbol{x}$, and then get its RGB value $\boldsymbol{y}_{RGB}$ by:

$$\boldsymbol{y}_{RGB} = (\boldsymbol{x} + 1)/2. \tag{1}$$

Then when two gaze labels are close, their colors should be close too. It can be seen that gaze is mixed in eye features, while in gaze features, similar gaze tends to gather together.

## References

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2, 3

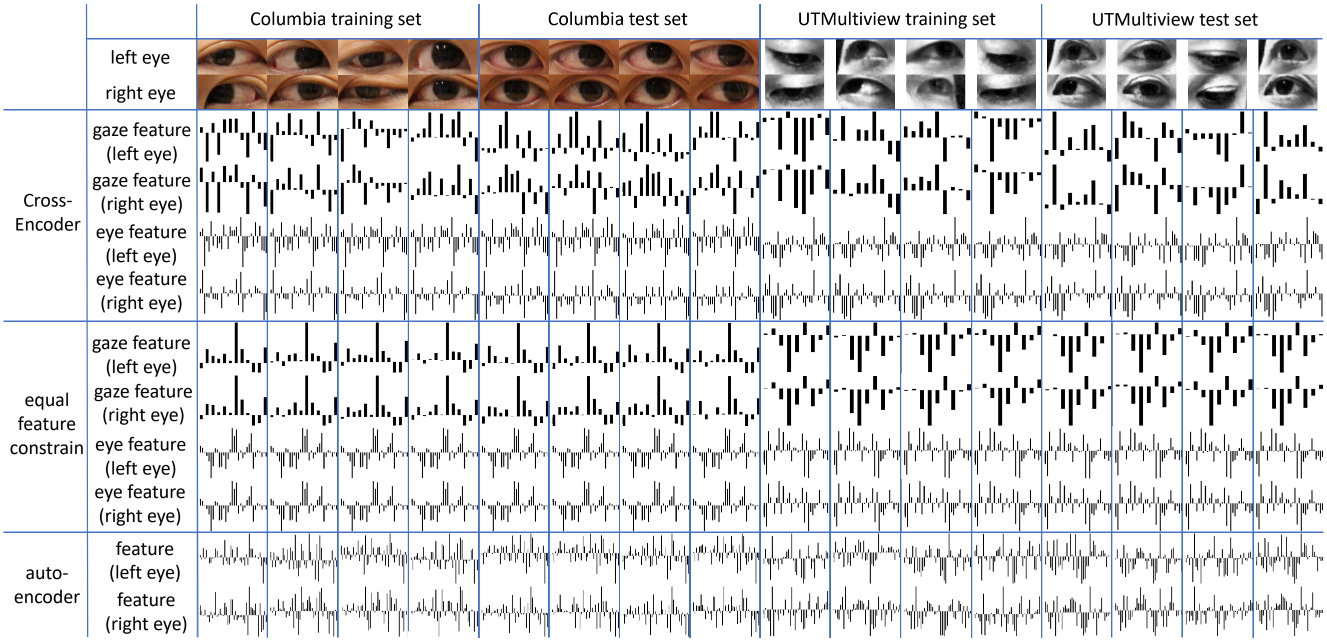
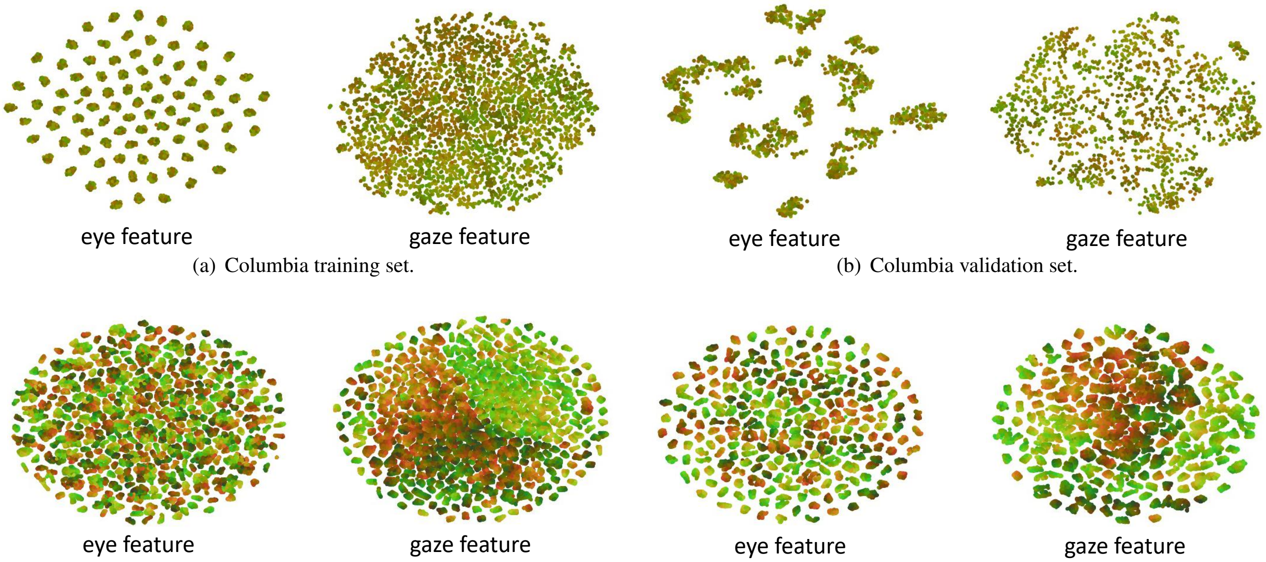


Figure 1. More examples of eye images from different datasets and their corresponding representations. Eye images in the same column are of the left and right eye from the same video frame. For each dataset, the four eye images in the same row are from the same person.



(a) Columbia training set.

(b) Columbia validation set.

(c) UTMultiview training set.

(d) UTMultiview validation set.

Figure 2. Visualization of Cross-Encoder-learned representation using t-SNE[1]. Each point corresponds to an eye's image and is colored by its gaze label. Better viewed in color.