Supplementary Material of
Unsupervised Representation Learning for Gaze Estimation

1. Image Resolution

The image resolution of Columbia Gaze and UTMultiview samples is 36*60, while the resolution of Eyediap is 60*75.

2. Network Architecture

The detailed architectures of the Gaze Representation Learning Network $\mathcal{G}_\phi$, the Global Alignment Network $\mathcal{A}_\psi$ and the Gaze Redirection Network $\mathcal{R}_\theta$ are illustrated in Fig. 1, Fig. 2 and Fig. 3 respectively. They are based on ResNet blocks. Because of the different image resolution (Eyediap input images are larger), the architectures employed to handle Eyediap samples are a bit different than those for the Columbia Gaze and UTMultiview datasets. They mainly differ in pooling operations, and have been mentioned in the figures. Note that there are ReLu activation functions between the layers, which are omitted in the figures.

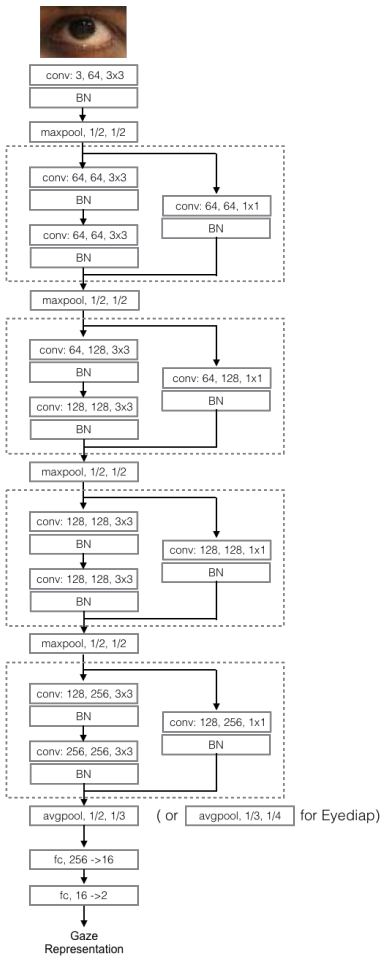

Figure 1. Gaze Representation Learning Network $\mathcal{G}_\phi$

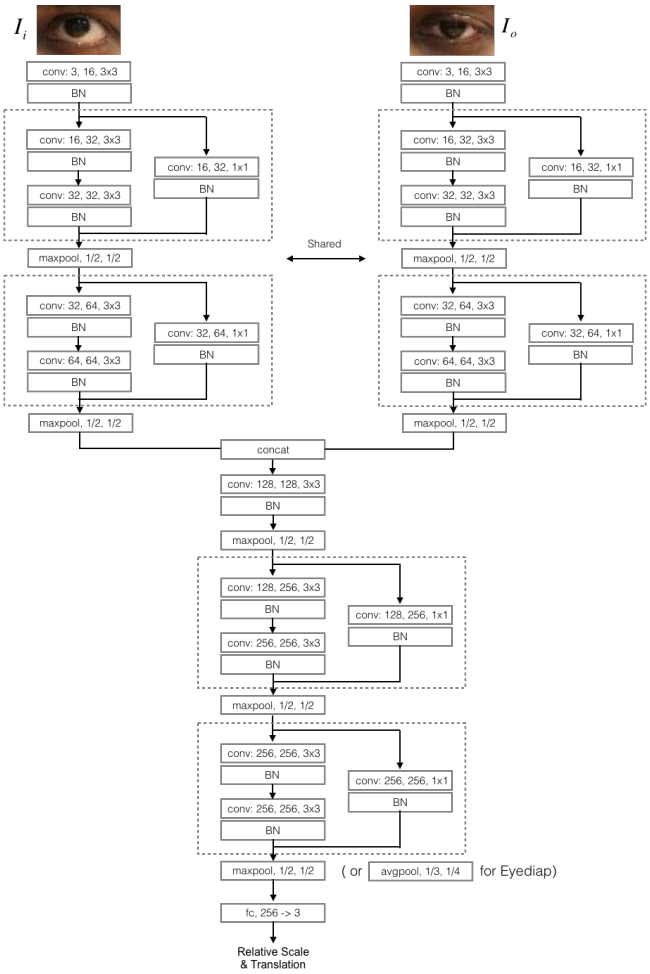

Figure 2. Global Alignment Network $\mathcal{A}_\psi$

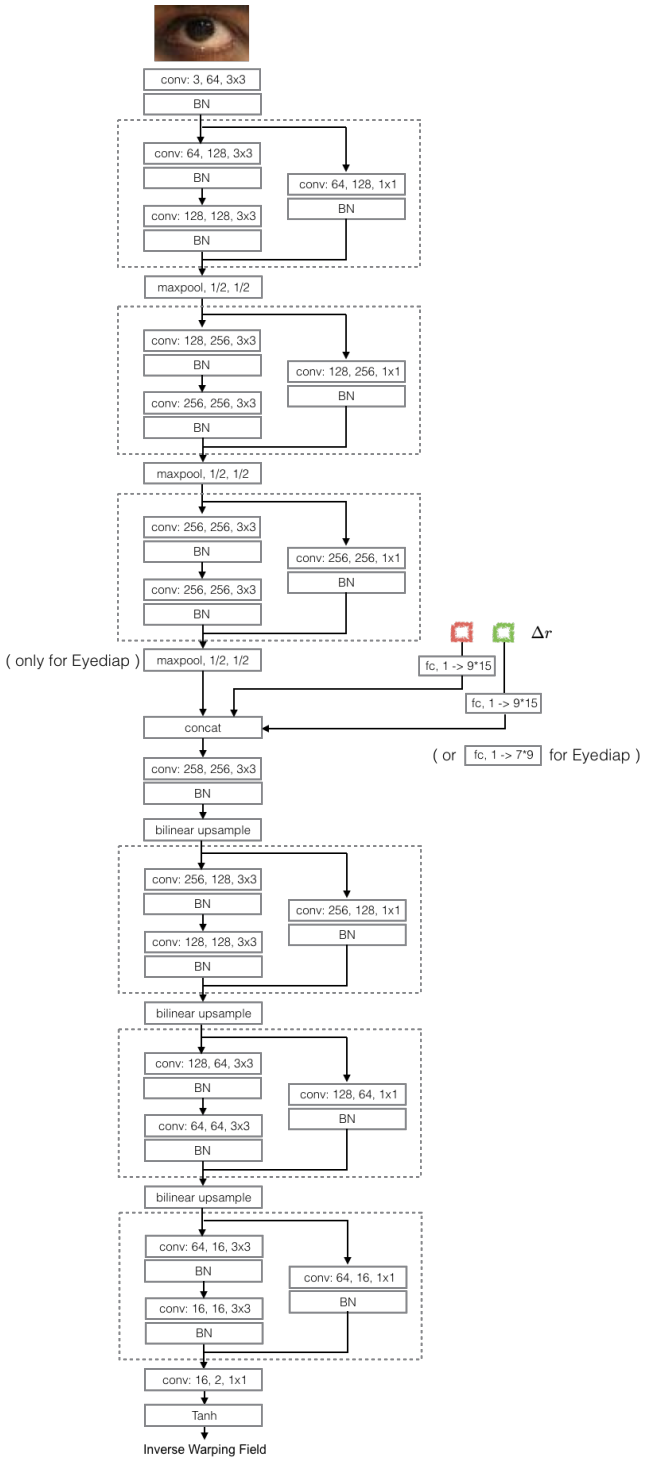

Figure 3. Gaze Redirection Network $\mathcal{R}_\theta$