The GazeCapture dataset is almost the largest gaze dataset in mobile device. We first conduct performance evaluation of our method on the GazeCapture dataset. We choose four methods for comparison on GazeCapture, which are iTracker, SAGE, TAT and AFF-Net. To the best of our knowledge, AFF-Net shows the SOTA performance on GazeCapture. We list the result in Table I. Our proposed method achieves 1.58 cm error on mobile phone captured images and 2.36 cm error on tablet captured images, outperforms SOTA methods on mobile phones. For mobile phone image test, the earliest iTracker has the highest error as 1.86 cm. SAGE and TAT has similar performance around 1.77 cm, improve about 5% from iTracker. AFF-Net achieves 1.62 cm error, outperforms these two methods significantly. Our proposed method achieves 1.58 cm, outperforms AFF-Net, and improves about 15% from iTracker. For the more challenging tablet image test, the error of iTracker is 2.81 cm. SAGE and TAT has similar performance around 2.69 cm. AFF-Net achieves 2.3 cm error, which is almost 0.39 cm lower than SAGE and TAT. Our proposed method achieves similar results to AFF-Net and the error is 2.36 cm, which is 0.33 cm lower than SAGE and TAT. As there are only about 15% images are from tablets, the results show that our proposed method can also relatively improve the performance of gaze estimation on tablets than most previous methods. These experiment results show that our proposed method has a clear advantage compare with other methods, especially on mobile phones.

To further demonstrate the advantage of our method, We conduct more experiments on the MPIIFaceGaze dataset. We calculate face and eye bounding boxes according to provided facial landmarks and convert the screen pixel coordinates of targets to physical distance. We choose iTracker and Spacial Weights CNN as the compared methods, since they both show outstanding performance in the 2D gaze position estimation task on MPIIFaceGaze. Meanwhile, since MPIIFaceGaze is popular used in 3D gaze direction estimation task, we also select RT-GENE as compared method for providing convinced comparison. The RT-GENE almost shows start-of-the-art performance in 3D gaze direction estimation task on MPIIFaceGaze. In order to provide more comprehensive comparison, we further convert the 2D gaze positions result estimated from our AFF-Net into 3D gaze directions according to provided camera-screen calibration matrix. As can be seen in Figure 5 and Table II, our method achieves the performance of 3.9 cm Euclidean error and 4.4 degree angular error, which significant performs better than other compared methods. Note that, the MPIIFaceGaze dataset is collected from laptop. This result demonstrate that our method also can perform well in the laptop.