The Benefits of Depth Information for Head-Mounted Gaze Estimation

STEFAN STOJANOV, Georgia Institute of Technology, USA
SACHIN TALATHI, Reality Labs Research, Meta, USA
ABHISHEK SHARMA, Reality Labs, Meta, USA

In this work, we investigate the hypothesis that adding 3D information of the periocular region to an end-to-end gaze-estimation network can improve gaze-estimation accuracy in the presence of slippage, which occurs quite commonly for head-mounted AR/VR devices. To this end, using UnityEyes we generate a simulated dataset with RGB and depth-maps of the eye with varying camera placement to simulate slippage artifacts. We generate different noise profiles for the depth-maps to simulate depth sensor noise artifacts. Using this data, we investigate the effects of different fusion techniques for combining image and depth information for gaze estimation. Our experiments show that under an attention-based fusion scheme, 3D information can significantly improve gaze-estimation and compensates well for slippage induced variability. Our finding supports augmenting 2D cameras with depth-sensors for the development of robust end-to-end appearance based gaze-estimation systems.

Additional Key Words and Phrases: gaze estimation, deep neural networks, slippage robust, fitment robust

ACM Reference Format:
Stefan Stojanov, Sachin Talathi, and Abhishek Sharma. 2022. The Benefits of Depth Information for Head-Mounted Gaze Estimation. In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22), June 8–11, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3517031.3529638

1 INTRODUCTION

Accurate human gaze estimation is essential for applications such as general human computer interaction [Lutteroth et al. 2015; Menges et al. 2017] or more specific applications such as augmented and virtual reality [Chu et al. 2020; Patney et al. 2016], biometrics [Lohr et al. 2020] and behavior analysis [Huang et al. 2016; Rogers et al. 2018; Slone et al. 2018]. It is especially important to support some of the critical AR/VR technologies such as foveated-rendering [Patney et al. 2016], telepresence [Chu et al. 2020; Lombardi et al. 2018], action anticipation and recognition [Bulling et al. 2009; Li et al. 2021], visual memory recall [Bulling and Roggen 2011], and visual search [Sattar et al. 2017]. Earlier gaze-estimation approaches are based on the geometric eye-model and rely on complete view of the eye and known camera and lighting positions w.r.t. the eye [Guestrin and Eizenman 2006b; Hansen and Ji 2010]. Recently, deep-learning appearance-based models, which directly regress the gaze-direction from the eye-image in some world-coordinate system, have received a lot of attention [Cheng et al. 2021; Ghosh et al. 2021] as an alternative direction.

Such deep-learning models are especially well suited for AR/VR gaze-estimation because the camera position is generally off-axis, below and off to the side of the eye (see Figure 1), resulting in a partial view of the eye, which renders model-based techniques less accurate [Guestrin and Eizenman 2006b]. Further, the power, form-factor and heat-dissipation constraints require minimal hardware components. Unfortunately current appearance-based gaze-estimation systems, including deep learning-based ones, are extremely sensitive to the variation in camera-placement (see Tab.1). Such variations are quite common for head-mounted-devices, or HMDs, due to vastly varying head and face structures, referred to as *fitment variability*. Fitment variability can be compensated by user-calibration [Guestrin and Eizenman 2006b], but it leads to poor user-experience. Moreover, the common use of HMDs—playing games, watching 360 scenes—involves head-motion that leads to HMD's *slippage*, which may not be fully captured in the data available to train appearance based eye-tracking models. . Therefore, despite the minimal assumptions, appearance-based gaze-estimation systems still have unresolved challenges and further research is warranted to make appearance based eye tracking models to various types of fitment and slippage induced variability.

In this work, we seek motivation from the aforementioned gaps in appearance-based gaze estimation approaches for HMDs and investigate the improvements offered by depth information. Our investigation focuses on tackling issues caused by variations in sensor-position, namely fitment and slippage variability. Moreover, we restrict our focus to deep-learning based methods owing to their great success [Cheng et al. 2021; Ghosh et al. 2021]. For HMD gaze-estimation, a depth sensor corresponding to the 2D input of a camera may encode valuable signal like the position of the camera relative to the periocular region, the shape of the eyelid and depending on the sensor type, and potentially the orientation of the iris plane. These factors are important for estimating gaze and are more easily resolved from depth information than images alone. As depth sensor technologies become increasingly smaller and more precise as a result of their continued development [Horaud et al. 2016], it is essential to understand the potential impact of their application for HMD gaze estimation.

Since such depth sensor technologies are currently unavailable, we leverage synthetic-data generated using a modified implementation of UnityEyes [Wood et al. 2016] that can yield depth-maps corresponding to the rendered images and segmentation maps for the skin, cornea, iris and pupil. We use these segmentation-maps are used to introduce region-specific noise for increased depth realism. We use a modified Residual CNN [He et al. 2016] and conduct extensive evaluations combining image and depth using early and late fusion, as well as using intermediate cross-modal attention mechanisms to understand the benefits of depth-maps towards slippage robustness. In summary, our contributions are as follows:

(1) A large synthetic dataset that simulates both fitment and slippage variability for 85 different synthetic identities with depth and image information.
(2) The first extensive study spanning multiple architecture and modality fusion approaches demonstrating the benefit of depth information for head-mounted gaze estimation.
(3) An analysis of the causes of degraded performance for deep networks in the presence of fitment and slippage variability and the improvements due to depth and their robustness to input noise.

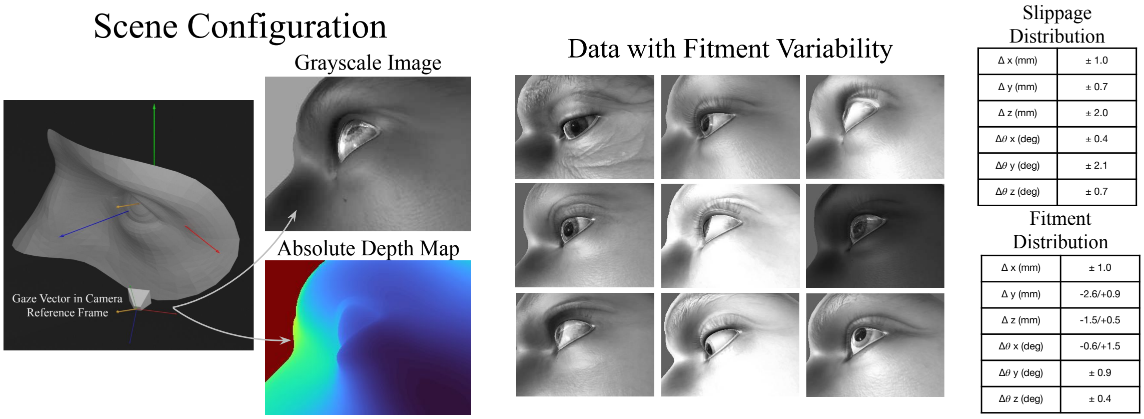
2 BACKGROUND AND RELATED WORK

An overview of gaze estimation techniques: Gaze-estimation approaches can be broadly categorized into 3D eye model-based and appearance based-systems [Kar and Corcoran 2017]. 3D model-based systems obtain a person-specific geometric model of the eye and use that to estimate gaze-direction, we refer the readers to [Hansen and Ji 2010] for a comprehensive review. Owing to the use of explicit geometric modeling, such systems are insensitive to the sensor-placement and require a user-calibration step to personalize the eye-model [Guestrin and Eizenman 2006a]. Our work falls in the appearance-based systems where the eye-image is passed through some feature-extractor and the gaze-direction or point-of-gaze is regressed, see [Jiang et al. 2019] for a review. Among appearance-based gaze-estimation methods, deep-learning systems have become immensely popular due to their powerful feature-learning ability from data, which does away with the requirement of carefully hand-crafted feature for regression, see [Cheng et al. 2021; Ghosh et al. 2021] for a comprehensive review.

Gaze-direction for appearance-based systems: Eye-model based approaches refer to the optical axis of the eye as the gaze direction and employ the known or estimated geometric transformation to project the optical axis in the desirable coordinate systems, often a display screen in the form of point-of-gaze, or intersect the optical axis with a 3D scene. Appearance-based approaches, on the other hand, regress for the gaze-direction directly in some external coordinate system. This is due to the lack of an intrinsic eye-model followed by a personal calibration step to account for the difference between visual and optical axis [Guestrin and Eizenman 2006b; Park et al. 2019]. The estimated gaze direction is then projected to the desired coordinate-system as explained above. It is important to note that the external coordinate system is defined either explicitly, through geometric transformations, or implicitly "anchored" to the sensor so that the gaze-direction is regressed w.r.t. the sensor's coordinate system. Therefore, the estimated gaze-direction is sensitive to the sensor's location and orientation w.r.t. the eye. This sensitivity can be especially problematic for head-mounted gaze-estimation systems that undergo frequent *slippage* from their original position due to head motion/orientation, skin/tissue deformation and frequent headset adjustments by the user [Santini et al. 2019]. Specifically for deep learning methods, our experiments show that the fitment and slippage variations can lead to dramatic reduction in gaze estimation accuracy, see Sec.4.

Slippage Robust Gaze Estimation: The aforementioned slippage problem has been addressed by some prior work. The method in [Santini et al. 2019] employed the pupil outline to obtain hand-crafted slippage-invariant features for gaze regression. Although, this achieves impressive improvements, its strong reliance on pupil visibility limits its applications. Similarly deep-learning approaches try to disentangle head motion and eyeball rotation to obtain sensor-placement robustness. An unsupervised framework to learn a low dimensional eye representation with gaze redirection was proposed in [Yu and Odobez 2020]. The Disentangling Transforming Encoder-Decoder, framework [Park et al. 2019] tries to disentangle identity, head pose and gaze direction and employs few-shot learning to personalize the identity and head pose sub-networks during calibration for each person. Unfortunately, it is still sensitive to head motion after calibration. The Self-Transforming Encoder-Decoder architecture [Zheng et al. 2020] approach employs a generative model to dis-entangle head and eye-ball rotation. Different from all past approaches, we propose to fuse depth-maps with RGB images to endow slippage robustness to an end-to-end deep-learning system. In order to assess the improvement offered from depth maps, we also simulate an off-axis gaze-camera dataset, using UnityEyes [Wood et al. 2016], with slippage for 85 synthetic identities. This dataset can serve as a benchmark to test similar algorithms for slippage robustness in the future.

Depth-map for Gaze-Estimation: While appearance-based gaze-estimation has seen a lot of work with 2D camera sensors, relatively less importance has been given to other modalities, such as depth maps. Thanks to the recent advances in the 3D sensing technologies, there exist 3D sensors that can fit into the form-factor and power budget of head-mounted VR systems. Therefore, exploration of such hybrid sensing modalities for gaze-estimation is a promising direction. The work in [Lian et al. 2019] leverages both depth maps and RGB images for remote gaze-estimation. Our work, on the other hand, focuses on a head-mounted gaze-estimation system and we focus on exploring the benefits of depth-maps for slippage robustness. Moreover, [Lian et al. 2019] uses a different fusion mechanism for merging the information from depth and RGB frames. Just like gaze-estimation, there is an abundance of prior research on fusion methods, which is impossible to cover due to space limitations, hence, we refer the reader to [Gao et al. 2020] for a comprehensive review of deep-learning based fusion approaches. Some of the most common techniques are feature concatenation/summation/multiplication at different layers of abstractions giving rise to early, mid and late fusion


Scene Configuration
Grayscale Image
Absolute Depth Map
Gaze Vector in Camera Reference Frame
Data with Fitment Variability

Slippage Distribution

Δ x (mm)	± 1.0
Δ y (mm)	± 0.7
Δ z (mm)	± 2.0
Δθ x (deg)	± 0.4
Δθ y (deg)	± 2.1
Δθ z (deg)	± 0.7

Fitment Distribution

Δ x (mm)	± 1.0
Δ y (mm)	-2.6/+0.9
Δ z (mm)	-1.5/+0.5
Δθ x (deg)	-0.6/+1.5
Δθ y (deg)	± 0.9
Δθ z (deg)	± 0.4

Fig. 1. **Left:** The 3D scene configuration used for generating grayscale images and depth maps. Our goal is to predict gaze in the camera reference frame. **Middle:** Images from different synthetic identities generated with fitment variability. Note how the eye position is different in each image. **Right** Sampling ranges for slippage and fitment distributions.

frameworks. We carry out thorough ablation studies to study the impact of different fusion-methods and also employ a novel Cross-Modal Attention module [Zhang et al. 2022] that shows much better results than simple fusion techniques.

3 APPROACH

There are two requirements to achieve our goal of investigating the effect of incorporating depth to 2D image data:

Paired Image and Depth Data: Currently there are no head mounted devices on the market which allow for such data collection. We therefore use a modified UnityEyes [Wood et al. 2016] implementation to generate images, absolute depth maps and corresponding 3D gaze direction in the camera coordinate system and simulate individual fitment variability and slippage variability.

Baseline Architecture and Fusion Technique: Directly applying a Residual CNN [He et al. 2016] to regress in-camera 3D gaze direction results in surprisingly high performance (Table 1). We aim to both understand the effect of absolute depth and to what extent end-to-end models with increasing realism can directly make use of this additional source of information. We investigate early fusion, late fusion and attention-based fusion.

3.1 Synthetic Data Generation

We modified an open-source periocular simulator, UnityEyes [Wood et al. 2016], to generate eye-images and depth-map with 3D gaze direction (eye-in-head spherical coordinate system), see visualization in Please refer to Fig. 1. Our concurrent submission [anonymous [n. d.]] [1] developed a means of generating different synthetic identities based on varying face/eye morphology, iris and skin textures. We used it to create 100 synthetic identities with different appearances, from which we form a 70/15/15 training/validation/test split to evaluate appearance based models' generalization ability across individuals.

Fitment and Slippage Variability: In addition to individual-specific appearance variability, we also define two distributions 1) a uniform fitment variability distribution from which we sample a fixed fitment transform (rotation and translation) for each identity and 2) a uniform slip variability distribution over another predefined nominal range.

[1] Please see Section 3.1, of the anonymized manuscript provided in the supplement for these details.

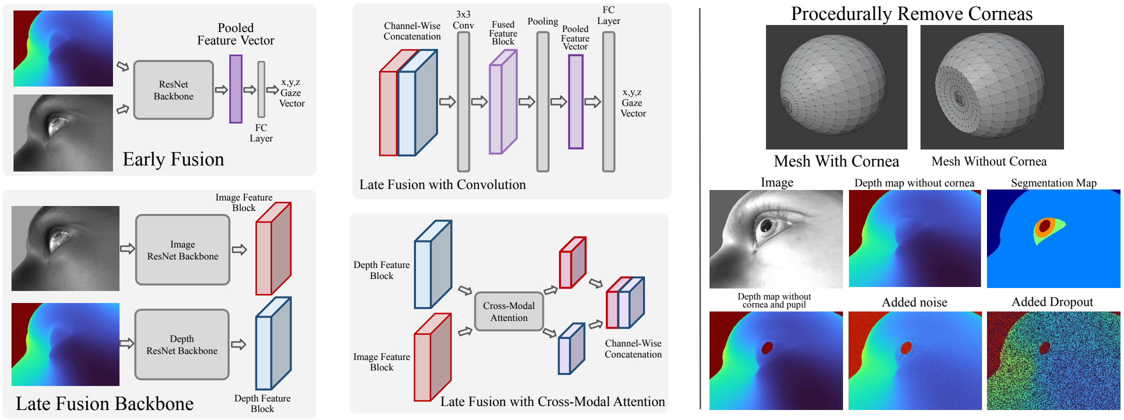


Fig. 2. **Left:** Early fusion and two late fusion architectures, convolutional late fusion and late fusion with cross modal attention. **Right** We generate more realistic depth information by removing the cornea and adding Gaussian noise and dropout to the depth map.

Although facial geometry is vital for fitment-estimation, UnityEye only simulates peri-ocular regions which isn’t sufficient for fitment-estimation, therefore, we use a uniform distribution. To generate the camera pose for one data sample from a specific individual, we first apply that individual’s fitment transform to a predetermined initial off-axis camera position that is constant, and then sample and apply another transform from the slip distribution.

Camera/Headset Reference Frame: AR/VR applications require the knowledge of gaze-direction relative to the content served by the headset. Naively, one could first estimate the absolute (eye-in-head spherical coordinate) eye pose, and then convert it to the HMD reference frame. However, it is generally not possible to find the geometric transform for this conversion, because (1) headsets significantly vary in their fitment across individuals due to facial structure (2) fitment changes across sessions and potentially slip during use. As the eye tracking camera is mounted on the headset itself, it therefore is a requirement of the AR/VR-HMD setting for the gaze direction to be predicted in the *camera coordinate frame* in a manner that generalizes across individuals and is robust to fitment and slippage.

Depth Map Generation: We generate depth map corresponding to the gray-scale images. However, these depth maps have no noise and are therefore unrealistic. For example, for time of flight-based depth sensors, the cornea depth cannot be resolved due to refraction and the pupil would not be visible at all due to light absorption. To improve the realism of our data, we add Gaussian noise with varying intensity for the skin, sclera, and iris and completely remove the pupil (we use noise with std. 0.1cm for the skin, std 0.2cm for the sclera and std 0.4cm for the iris). Further, as sensor measurements can be lossy, we also dropout each pixel with 0.5 probability. In addition, prior to generating the depth maps we modify the eyeball mesh to remove the cornea. For each individual we generate gaze samples at 0.5 increments in the $[-25°, 25°]$ range (relative to the eye’s center of rotation), resulting in 10,000 samples. For added realism, each sample is generated with random light intensity and pupil size sampled from a Gaussian distribution.

3.2 Appearance Based Gaze Estimation Baselines

We adopt a CNN-based model as our baseline architecture. We choose a ResNet18 [He et al. 2016], followed by a fully connected layer that outputs a normalized vector $\mathbf{u} = [x, y, z]$ indicating the gaze direction in the camera coordinate

Table 1. Early fusion of depth and image information improves gaze estimation angular error *in degrees*. **NC**: No Cornea, **NC+D** No Cornea + Dropout, **NC+D+N:** No Cornea + Dropout + Noise. First, and most importantly, we note that the fitment and slippage variability significantly reduces the performance as compared to the fixed set. Second, We find that the increased variability of the Fit+Slip improves the performance of the model. Last, increasing the depth-realism significantly reduces the benefits of depth for simple early fusion.

Approach/Dataset	*Fixed*	Fit	Fit+Slip	Fit NC	Fit + Slip NC	Fit NC+D	Fit + Slip NC+D	Fit NC+D+N	Fit + Slip NC+D+N
Image Only	*0.31*	1.66	1.19	1.66	1.19	1.66	1.19	1.66	1.19
Image + Depth (Early Fusion)	*N/A*	**1.28**	**0.77**	1.37	0.74	1.43	0.91	1.41	1.12

system. The simplest means of combining depth and grayscale image information we investigate is by concatenating them along the channel dimension as inputs to the CNN. We now describe the late fusion approaches.

Late fusion of image and depth information: We investigate two techniques for late fusion, using two separate backbone networks for feature learning, illustrated in Figure 2. Our first approach consists of channel-wise concatenation of the image and depth feature blocks, followed by a 3x3 convolution and standard average pooling to obtain a final feature. Our second approach is to apply a state-of-the-art transformer-based cross-modal attention block [Zhang et al. 2022], or CMA for short, to fuse the information between the depth and image features which outputs two feature blocks. These are then concatenated and passed through a convolution as described in the prior convolution-based fusion.

4 EXPERIMENTS

This section describes the implementation details of our baseline methods, followed by our empirical findings about the utility of depth information for gaze estimation in the presence of fitment and slippage variability. Performance is reported as angular error between the ground truth and predicted gaze vectors in the camera reference frame.

Implementation Details We render grayscale images in UnityEyes with a resolution of 240×320. We generate three distinct datasets: (1) **Fixed** where the camera location is fixed for all the images across training and testing sets to serve as the ideal dataset and uses only image as input with the baseline neural-network to yield **0.31 degrees** mean gaze-error. (2) **Fit** where a single fitment per identity is sampled from twice the nominal range of fitment variability. (3) **Fit+Slip** where an additional slippage transformation is added to each sample from twice the nominal slippage range. For the CNN we use a ResNet18 [He et al. 2016] backbone with half the channel width, which we train with stochastic gradient descent algorithm using cosine decay and an initial learning rate of 0.05. For the loss we use an L_1 loss between the predicted gaze direction and ground truth gaze.

4.1 Early Fusion of Depth Leads to Improvements but is Brittle

We present results for our investigation of adding depth with early fusion in Table 1. First, and most importantly, we note that the fitment and slippage variability increases the gaze-error vs. the fixed set by 5.35x and 3.83x, respectively. This observation supports our claim that deep-learning models for gaze-estimation are extremely sensitive to fitment and slippage variability. Second, we find that simple early fusion yields 23% (1.66 to 1.28) and 35% (1.19 to 0.77) relative improvement for Fit and Fit+Slip, respectively. We further study the effect of depth in this setting by removing the cornea (NC), adding dropout (D) and Gaussian noise (N) and present the results in Table 1. Gaze-error successively

Table 2. Late fusion of depth and image information improves gaze estimation angular error where **NC+D+N** indicates No Cornea + Dropout + Noise. We find that late fusion with cross-modal attention is essential for obtaining significant performance improvements in the noisy settings as compared with early fusion. Further, removing the cross-modal attention (last row) reduces performance.

Approach / Dataset	Fit NC+D+N	Fit + Slip NC+D+N
Image Only	1.66	1.19
Image + Depth (Early Fusion)	1.41	1.12
Image + Depth (Late Fusion Conv + CMA)	**1.23**	**0.87**
Image + Depth (Late Fusion Conv)	1.53	-

Table 3. A simple calibration fitment using 9 points for testing identities results in significant performance improvement. Interestingly, this finding holds for data with slippage variability as well but by a smaller margin. This indicates that appearance-based CNNs overfit to fitments in the training set.

Dataset / Approach	Image Only Uncalibrated	Image Only Calibrated	Image + Depth Uncalibrated	Image + Depth Calibrated
Fit	1.67	**0.99**	1.23	**0.73**
Fit + Slip	1.194	**0.935**	0.86	**0.64**

increased with added realism and it reflects that the naive early fusion model would likely not yield benefits in real world scenarios. Interestingly, we find higher performance with Fit+Slip training dataset vs. Fit only training dataset. It's expected because the Fit dataset has only 70 camera views for training, one per identity. On the other hand, Fit+Slip dataset exposes the network with $70 \times 10000 = 700K$ different camera-views, which allows the network to build invariance against camera-view changes. This potentially means that future real data collection protocols should include sessions with multiple headset placements for each individual. We further study this finding in Section 4.3.

4.2 Late Fusion for Depth Information Works with Increased realism

We present results for this setting in Table 2. We find that late fusion with CMA works significantly better in the realistic setting with removed cornea, dropout and Gaussian noise, resulting in 26% relative improvement compared with 15% for early fusion for Fit. Similarly, for Fit + Slip we find a 27% relative improvement compared with 6% for early fusion. Thus, we conclude that CMA is critical for fusing depth with appearance under more realistic sensor conditions. It's due CMA's attention mechanism for fusing multi-modal information, which affords adaptive selection of the depth-region to tackle sensor noise while fusing it with appearance.

4.3 Deep-Learning Models Can Easily Over-fit to Fitment/Camera Pose

In Sections 4.1 and 4.2 we observed that the both image only and image + depth models perform better in the Fit + Slip setting, which has significantly higher variability for both training and testing data. We hypothesize that performance in the Fit setting is worse as a result of over-fitting to the individual fitments in the training set.

We test this hypothesis using the Fit dataset for training and use a simple linear calibration scheme [Guestrin and Eizenman 2006b] to learn an angular-offset for each individual. We keep the same training procedure, however prior to testing, we randomly draw 9 samples and their ground truth gaze direction and perform a fitment calibration procedure for each identity. The procedure is as follows: (1) Generate predictions for these 9 samples and convert the

Image Only

Test data → / Train data ↓	Uncalibrated Fit	Uncalibrated Fit + Slip	Calibrated Fit	Calibrated Fit + Slip
Fit	1.66	2.20	1.03	1.68
Fit + Slip	1.15	1.19	**0.88**	**1.02**

Image + Depth

Test data → / Train data ↓	Uncalibrated Fit	Uncalibrated Fit + Slip	Calibrated Fit	Calibrated Fit + Slip
Fit	1.28	1.96	0.74	1.71
Fit + Slip	0.63	0.64	**0.45**	**0.51**

Fig. 3. Evaluation all combinations of training with Fit and testing with Fit + Slip and vice versa for both Image Only and Image + Depth models. We find that data with the highest variability results in the highest generalization, and despite high variability simple fitment calibration still leads to improvement.

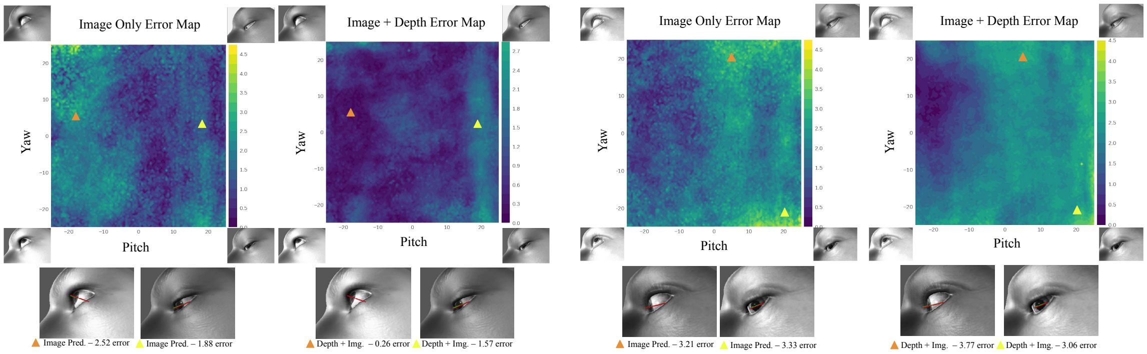


Fig. 4. Visualization of gaze error across all possible gaze directions. We observe that adding depth reduces the error across the board.

3D gaze vector into x_p-yaw and y_p-pitch angles (2) optimize for linear parameters a_1, a_2, b_1, b_2 using the equations $x_{gt} = a_1 x_p + b_1$ and $y_{gt} = a_2 y_p + b_2$ and the grount truth gaze directions x_{gt}, y_{gt} (3) Apply these learned parameters to all predictions for the identity. If the model does in fact overfit to the fitment in the training set, this simple calibration should significantly improve performance by compensating for how the novel identity is different from the training fitments. Our findings (see Table 3) provide evidence confirming our hypothesis.

4.4 Further Analysis

Since practically it is very difficult to collect ground truth data with slippage, a natural question is: To what extent is depth information useful when training models on data just with fitment variability, but testing in cases where there is both fitment and slippage? We find that in the most challenging, un-calibrated scenario using image and depth information results in improved generalization to novel identity testing data with fitment and slippage, while training is only done with different fitments. These results are provided in Figure 3. In addition, we visualize the distribution of gaze errors in Figure 4. We observe that adding depth information leads to improvements in error across the board, and that the most significant impact is obtained in cases where the iris and pupil are not very visible.

5 CONCLUSION

In this work we use synthetic data to investigate the utility of depth information for estimating gaze in AR/VR-HMD applications in the presence of fitment and slippage variability. We showed that with CNN-based fusion techniques, in particular the cross-modal attention-based fusion, adding depth information improves generalization to novel identities in the presence of fitment and slippage variability.

REFERENCES

anonymous. [n. d.]. Multi-Rate Sensor Fusion for Unconstrained Near-Eye Gaze Estimation.

Andreas Bulling and Daniel Roggen. 2011. Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th international conference on Ubiquitous computing*. 455–464.

Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. 2009. Eye movement analysis for activity recognition. In *Proceedings of the 11th international conference on Ubiquitous computing*. 41–50.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2021. Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. *CoRR* abs/2104.12668 (2021).

Hang Chu, Shugao Ma, Fernando De la Torre, Sanja Fidler, and Yaser Sheikh. 2020. Expressive telepresence via modular codec avatars. In *European Conference on Computer Vision*. Springer, 330–345.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation* 32, 5 (2020), 829–864.

Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. 2021. Automatic Gaze Analysis: A Survey of Deep Learning based Approaches. *CoRR* abs/2108.05479 (2021).

Elias Guestrin and Moshe Eizenman. 2006a. General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *Biomedical Engineering, IEEE Transactions on* 53 (07 2006), 1124 – 1133.

Elias Daniel Guestrin and Moshe Eizenman. 2006b. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering* 53, 6 (2006), 1124–1133.

Dan Witzner Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 478–500.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. 2016. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications* 27, 7 (2016), 1005–1020.

Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2016. Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the 24th ACM international conference on Multimedia*. 1395–1404.

Jiaqi Jiang, Xiaolong Zhou, Sixian Chan, and Shengyong Chen. 2019. Appearance-Based Gaze Tracking: A Brief Review. In *International Conference on Intelligent Robotics and Applications*. Springer, 629–640.

Anuradha Kar and Peter Corcoran. 2017. A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms. *IEEE Access* 5 (2017), 16495–16519.

Yin Li, Miao Liu, and James Rehg. 2021. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

Dongze Lian, Ziheng Zhang, Weixin Luo, Lina Hu, Minye Wu, Zechao Li, Jingyi Yu, and Shenghua Gao. 2019. RGBD Based Gaze Estimation via Multi-Task CNN. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (Jul. 2019), 2488–2495.

Dillon J Lohr, Samantha Aziz, and Oleg Komogortsev. 2020. Eye movement biometrics using a new dataset collected in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*. 1–3.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

Christof Lutteroth, Moiz Penkar, and Gerald Weber. 2015. Gaze vs. mouse: A fast and accurate gaze-only click alternative. In *Proceedings of the 28th annual ACM symposium on user interface software & technology*. 385–394.

Raphael Menges, Chandan Kumar, Daniel Müller, and Korok Sengupta. 2017. Gazetheweb: A gaze-controlled web browser. In *Proceedings of the 14th International Web for All Conference*. 1–2.

Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. 2019. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9368–9377.

Anjul Patney, Joohwan Kim, Marco Salvi, Anton Kaplanyan, Chris Wyman, Nir Benty, Aaron Lefohn, and David Luebke. 2016. Perceptually-based foveated virtual reality. In *ACM SIGGRAPH 2016 Emerging Technologies*. 1–2.

Shane L Rogers, Craig P Speelman, Oliver Guidetti, and Melissa Longmuir. 2018. Using dual eye tracking to uncover personal gaze patterns during social interaction. *Scientific reports* 8, 1 (2018), 1–9.

Thiago Santini, Diederick C Niehorster, and Enkelejda Kasneci. 2019. Get a grip: Slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*. 1–10.

Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2017. Visual decoding of targets during visual search from human eye fixations. *arXiv preprint arXiv:1706.05993* (2017).

Lauren K Slone, Drew H Abney, Jeremy I Borjon, Chi-hsin Chen, John M Franchak, Daniel Pearcy, Catalina Suarez-Rivera, Tian Linger Xu, Yayun Zhang, Linda B Smith, et al. 2018. Gaze in action: Head-mounted eye tracking of children's dynamic visual attention during naturalistic behavior. *Journal of visualized experiments: JoVE* 141 (2018).

Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. 131–138.

Yu Yu and Jean-Marc Odobez. 2020. Unsupervised Representation Learning for Gaze Estimation. arXiv:1911.06939 [cs.CV]

Zehua Zhang, David Crandall, Michael Proulx, Sachin S. Talathi, and Abhishek Sharma. 2022. Can Gaze. In *ETRA*.

Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. 2020. Self-Learning Transformations for Improving Gaze and Head Redirection. arXiv:2010.12307 [cs.CV]