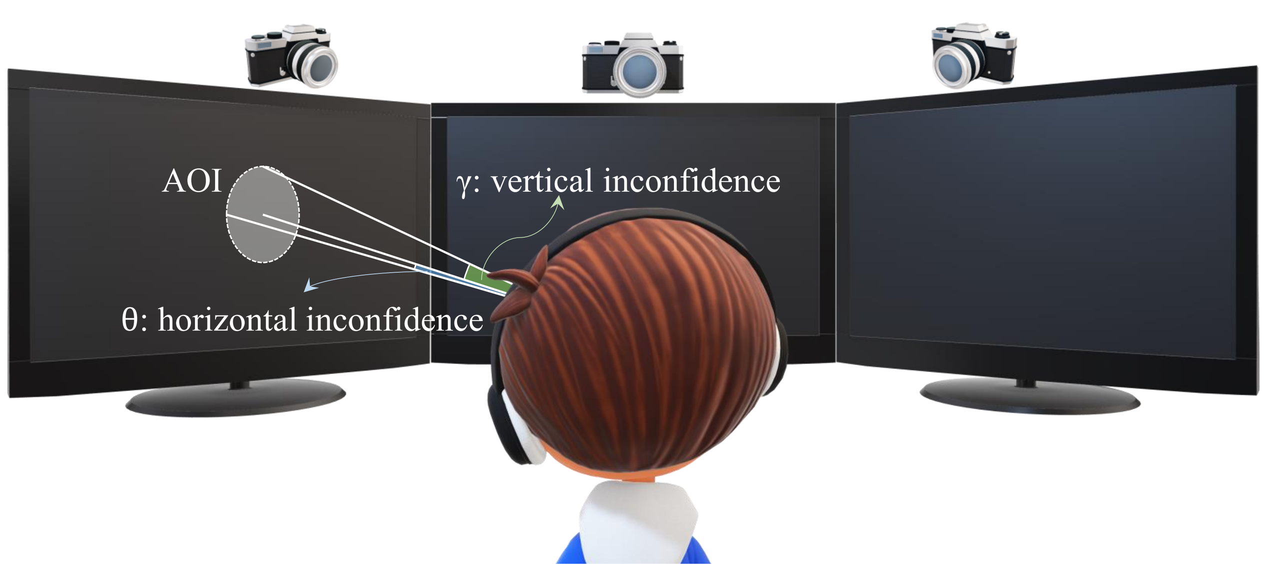
Generalizable Gaze and Gaze Zone Estimation through Variance and Invariance Learning

Xuchen Gong

xuchen.gong@duke.edu

Data Science | Class of 2022

昆山杜克大学 DUKE KUNSHAN UNIVERSITY



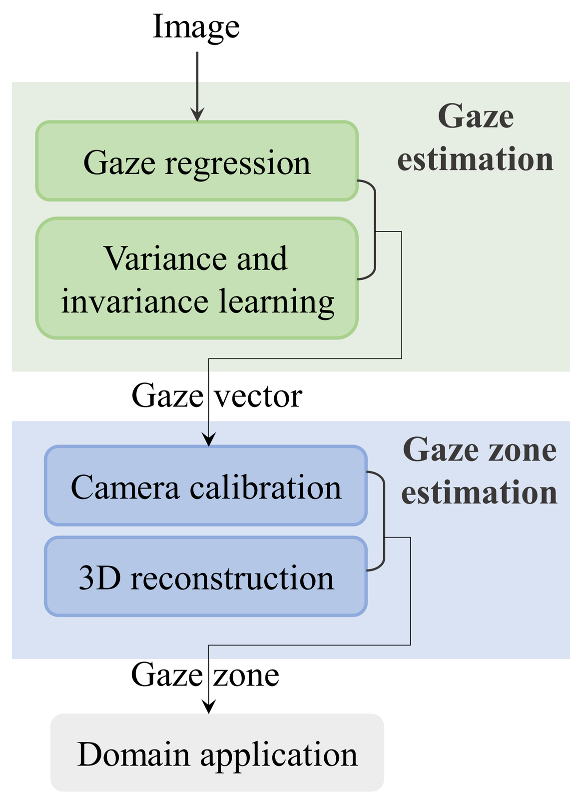


Figure 1: Our gaze and gaze zone estimation workflow. **Upper**: The estimated gaze and its horizontal and vertical angular error are used to predict the area of interest (AOI). **Right**: The estimated gaze and gaze zone can be applied to shopping scenarios, psychological research, and medical treatment.

Abstract

Gaze and gaze zone estimation have wide applications in VR/AR and social robotics for healthcare, medical treatment, and education. For example, a social robot can infer the users' intention through their gaze area; a company can infer the products' popularity through the customers' attention; cars can provide warning prompts when the drivers are not focusing.

However, gaze estimation algorithms' performance is suppressed by their sensitivity to different illuminations, subject identities, and viewing angles. Moreover, the performance of the models trained on datasets with different labeling methods is hard to be compared. Therefore, we study these two obstacles to the real-world application of gaze estimation algorithms.

1) To enhance robustness, we propose a variance and invariance learning framework for generalizable gaze estimation, whose effectiveness is evaluated by the models' angular error on the public dataset ETH-XGaze. 2) To enable model comparison, we propose a multi-view multi-screen 3D gaze reconstruction system, where three screens, three cameras, and the subject's gaze are visualized in one world coordinate system. Because of this system and our collected test videos, a model can be accessed quantitatively with our presented gaze zone error and qualitatively through visualization.

Related Work

- **Gaze estimation with domain adaptation.** To enhance the algorithms' robustness and generalizability to different identities, [8] adds an adversarial component into a CNN-based gaze estimator to learn features that can generalize appearance and pose variations. Similarly, to improve cross-domain performance in gaze estimation, [9] proposes an unsupervised domain generalization method to eliminate gaze-irrelevant features such as illumination and identity through gaze feature purification.
- **Multi-task learning.** [1] concatenated the head pose feature with the gaze feature to help estimate the gaze angle; OpenPose [7] employs a two-branch CNN to jointly predict confidence maps for body part detection and part affinity fields for parts association.
- **Invariant feature learning.** Most researchers apply data augmentation to make algorithms generalizable through "inclusion," while some "exclude" nuisance factors from the learned feature representations [2, 3, 4] through adversarial learning.
- **Gaze zone estimation.** Most studies focus on specific application scenarios, such as gaming platforms, website design, etc., which are not generally applicable and helpful for gaze estimation evaluation.

Gaze Estimation

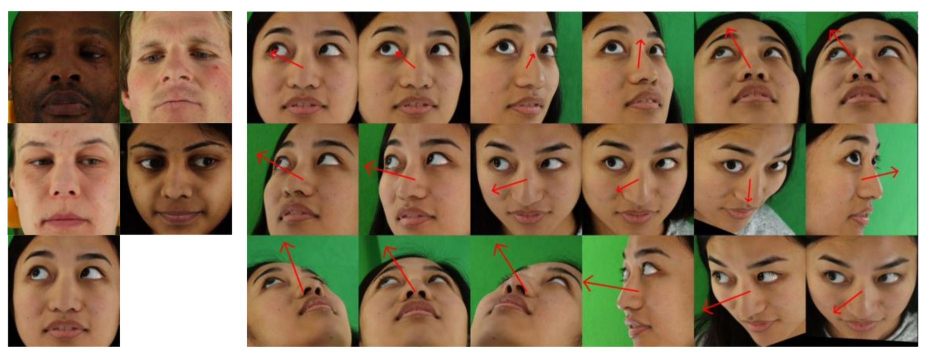


Figure 2: Example samples of the ETH-XGaze dataset. **Left:** The five selected diverse subjects for validation. **Right:** One gaze captured by 18 camera angles.

Definition	Symbol
# samples	S
# classes for feature h	K_h
# variant/invariant features	M/N
Binary indicator if class k is the correct classification	y_k

Table 1: Notations for gaze estimation.

$$L_v = -\sum_{h=1}^{M}\sum_{k=0}^{K_h} y_k \, log\left(D_h(E(x))\right)$$

$$L_{inv} = -\sum_{h=M+1}^{M+N}\sum_{k=0}^{K_h} \frac{1}{K_h} \, log\left(D_h(E(x))\right)$$

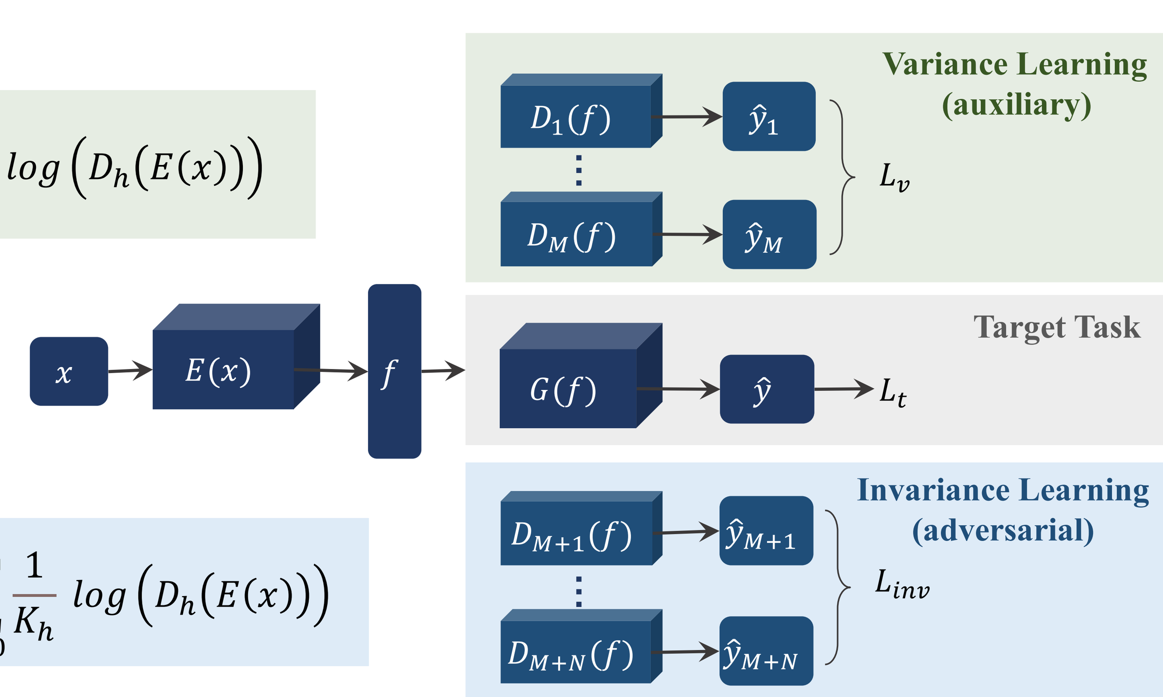


Figure 3: The generic variance and invariance learning framework.

$$L_{D_p} = -\sum_{k=0}^{K_h} y_k \, log\left(D_h(E(x))\right) \qquad L_E = L_t + \alpha L_v + \beta L_{inv}$$

Experiments

Table 2: Ablation study on ETH-XGaze.

Method	Pre-trained	Angular Error (°)
Baseline (ours)	ImageNet	4.5
+ variance (camera angle)	Baseline (ours)	4.4
+ invariance (identity)	Baseline (ours)	4.4
+ both	Baseline (ours)	**4.3**
ETH-XGaze's baseline [11]	ImageNet	4.5
SwAT [10]	VGG-Face	4.4

$$L_E = \frac{1}{S}\sum_{i=0}^{S}\left[\left\|G(E(x_i)) - g_i\right\|_1 - \alpha\sum_{k=0}^{17} y_{i,k}\log\left(D_{cam}(E(x_i))\right) + \beta\sum_{k=0}^{74} y_{i,k}\, log\left(D_{idty}(E(x_i))\right)\right]$$

Gaze Zone Estimation

Environment Building

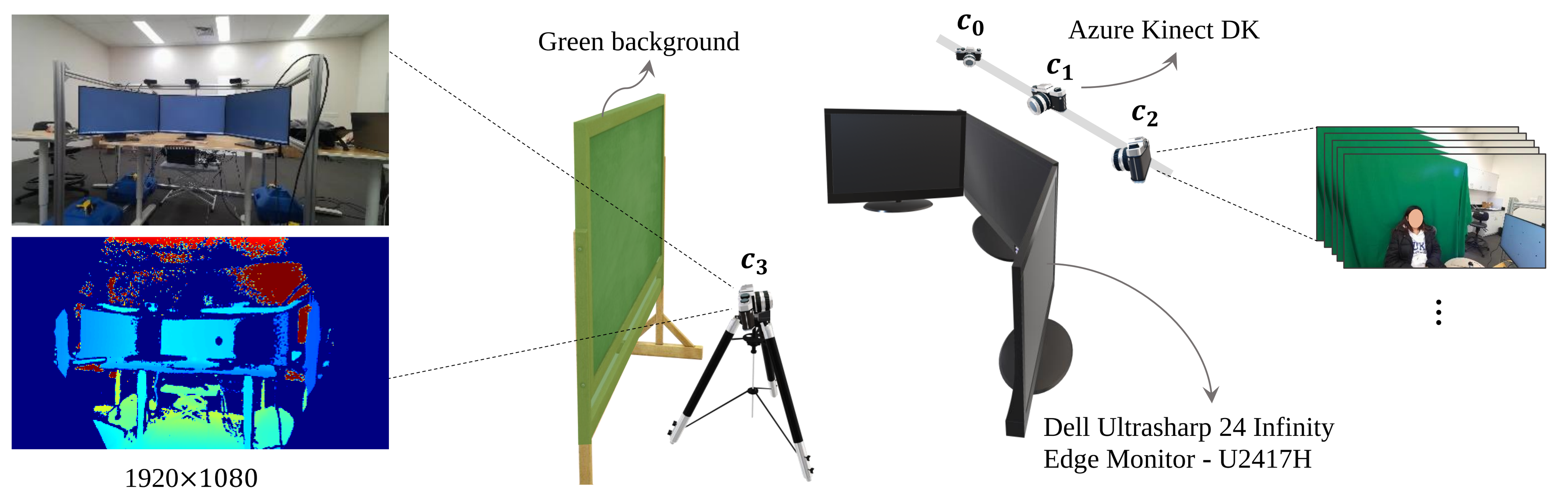


Figure 4: The test video capture environment. **Left:** An RGB and Depth frame of three screens taken by c_3. **Right:** Layout of three screens, three cameras, and the green background.

3D Reconstruction

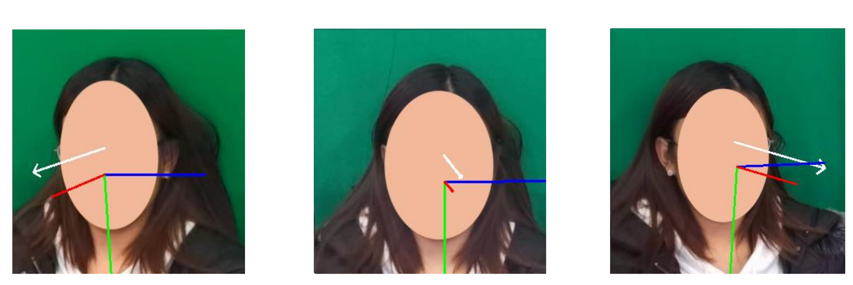
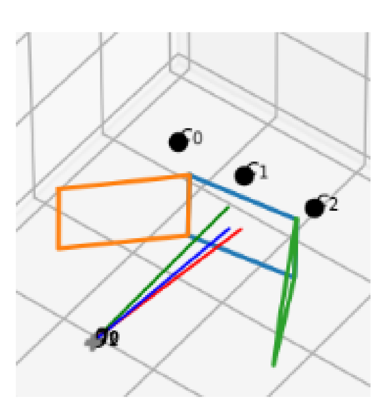


Figure 5: 3D reconstruction. The above three images are looking at p_{11} and the following vectors are looking at p_{13} (the middle point on the middle screen), with an error of 7-8 centimeters.



$$c_i\text{to}c_j = \begin{bmatrix} R_j & t_j \\ 0 & 1 \end{bmatrix}\begin{bmatrix} R_i^T & -R_i^T t_i \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} g_{x_w} \\ g_{y_w} \\ g_{z_w} \end{bmatrix} = R^{i\to w}\begin{bmatrix} g_{x_i} \\ g_{y_i} \\ g_{z_i} \end{bmatrix}$$

Definition	Symbol
Camera 0, 1, 2, 3	c_0, c_1, c_2, c_3
27 points on the screens	$p_0, \dots, p_{26}$
Gaze vector at c_i's camera coordinate system	$[g_{x_i}, g_{y_i}, g_{z_i}]^T$
Gaze vector at the world coordinate system	$[g_{x_w}, g_{y_w}, g_{z_w}]^T$
eye center at the world coordinate system	$[e_{x_w}, e_{y_w}, e_{z_w}]^T$
A matrix that transforms vectors at c_i's camera coordinate system to those at c_j's	$c_i\text{to}c_j$
Rotation matrix in $c_i\text{to}c_j$	$R^{i\to j}$
Point j's ground truth location in the world	$[p_{j_{x_w}}, p_{j_{y_w}}, p_{j_{z_w}}]^T$
Point j's predicted gaze point in the world coordinate system mapped from c_i's camera coordinate system	$[{p_j}'_{x_{w\leftarrow i}}, {p_j}'_{y_{w\leftarrow i}}, {p_j}'_{z_{w\leftarrow i}}]^T$

Table 3: Notations for gaze zone estimation.

Evaluation metric: Gaze Zone Error

$$GZE = \frac{1}{81}\sum_{i=0}^{2}\sum_{j=0}^{26}\left\|\begin{bmatrix} p_{j_{x_w}} \\ p_{j_{y_w}} \\ p_{j_{z_w}} \end{bmatrix} - \begin{bmatrix} {p_j}'_{x_{w\leftarrow i}} \\ {p_j}'_{y_{w\leftarrow i}} \\ {p_j}'_{z_{w\leftarrow i}} \end{bmatrix}\right\|_2$$

Significance: Different gaze datasets employ different labeling mechanisms, where some label the gaze vectors originating from eye centers while others from face centers, making angular errors of the algorithms incomparable. Therefore, it is valuable to transfer the gaze vectors to gaze points in a unified coordinate system and utilize gaze zone error (GZE) as a universal evaluation metric, which helps compare the algorithms trained on different datasets.

Further Work

To prove our proposed variance and invariance learning framework to be generally applicable to different tasks:

Invariance to inherent/synthetic nuisance factors. The Extended Yale B dataset [5] is an identity classification dataset with 38 subjects under 5 lighting conditions. Alternatively, we can apply data augmentation to generate the expanded MNIST [6], control the distribution of invariant factors, and apply our framework to the generated training samples.

Domain generalization. Domain gaps essentially come from invariant factors (lighting conditions, identities, background, etc.). Interpreting domain generalization as tasks where the training and testing set come from an exceptionally diverse dataset, we expect our framework to enhance performance in domain generalization.

Acknowledgement

This research is supervised by Prof. Ming Li and funded by Speech and Multimodal Intelligent Information Processing (SMIIP) Lab.

Conclusion

- We propose incorporating our prior knowledge of the variant and invariant features in a domain and a dataset **explicitly** during training, whose effectiveness in gaze estimation is demonstrated through our enhanced performance on ETH-XGaze.
- We present a multi-screen and multi-view gaze zone estimation system that 3D reconstructs the gaze vectors for visualization and model evaluation. It introduces gaze zone error, an evaluation metric for comparing gaze estimation algorithms that are trained on datasets with different labeling methods.

References

[1] Wang, Zhecan, et al. "Learning to detect head movement in unconstrained remote gaze estimation in the wild." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1):2096–2030, 2016.

[3] Yujia Li, Kevin Swersky, and Richard Zemel. Learning unbiased features. arXiv preprint arXiv:1412.5244, 2014.

[4] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zeme. The variational fair autoencoder. In Proceedings of International Conference on Learning Representations, 2016.

[5] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE transactions on pattern analysis and machine intelligence, 23(6):643–660, 2001.

[6] Jaiswal, A., Wu, R. Y., Abd-Almageed, W., & Natarajan, P. "Unsupervised adversarial invariance." Advances in neural information processing systems 31 (2018).

[7] Cao, Z., Simon, T., Wei, S. E., Sheikh, Y. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Wang, K., Zhao, R., Su, H., Ji, Q. "Generalizing eye tracking with bayesian adversarial learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[9] Cheng, Yihua, Yiwei Bao, and Feng Lu. "Puregaze: Purifying gaze feature for generalizable gaze estimation." arXiv preprint arXiv:2103.13173 (2021).

[10] Farkhondeh, A., Palmero, C., Scardapane, S., Escalera, S. (2022). Towards Self-Supervised Gaze Estimation. arXiv preprint arXiv:2203.10974.

[11] Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O. (2020, August). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In European Conference on Computer Vision (pp. 365-381). Springer, Cham.