GazeHELL: Gaze Estimation with Hybrid Encoders and Localised Losses with weighing

Shubham Dokania
shubham.dokania@mercedes-benz.com
Vasudev Singh
vasudev.singh@mercedes-benz.com
Shuaib Ahmed
shuaib.ahmed@mercedes-benz.com

Merdeces-Benz Research & Development India
Bangalore, India

Abstract

In the pursuit of robust eye gaze estimation, traditional approaches often grapple with the limitations of either spatial granularity or model interpretability. This paper introduces a dual-architecture framework that synergizes the strengths of Vision Transformers (ViT) and convolutional networks to enhance gaze estimation accuracy and reliability. We also propose two novel loss functions to refine our predictions: (1) a differentiable heatmap-based 2D MSE loss that transforms gaze vectors into a spatial heatmap enhancing the model's ability to localize gaze with high precision, and (2) a Fourier encoding loss that leverages high-dimensional Fourier features to capture complex spatial relationships more effectively. Additionally, we incorporate auxiliary uncertainty-based task weighing into our losses to provide a measure of confidence alongside gaze estimates, aiming to improve predictions dynamically during training. Our experimental results on MPIIGaze and RT-GENE datasets demonstrate significant improvements over existing methods and establishes a new state-of-the-art benchmark on both, with upto **3%** improvement in the respective datasets. This work not only advances the field of eye gaze estimation but also opens new avenues for applying advanced vision techniques in human-computer interaction and beyond.

1 Introduction

Eye gaze is a fundamental non-verbal communication cue, encapsulating a wealth of insights about human intent and attention. This valuable information has been pivotal in shaping numerous applications across diverse domains, including human-computer interaction [18, 27, 33], driver monitoring systems, reading analysis [3], dyslexia screening, and the rapidly expanding field of augmented reality [22]. Given the broad applicability and critical importance of these applications, achieving accurate eye gaze estimation is paramount.

One prominent approach to estimating human gaze is through appearance-based methods, which utilize visual perception to infer gaze direction from camera-captured images. Traditionally, deep networks leveraging convolutional architectures have been the dominant

technique, as evidenced by the work of [19]. However, recent advancements in transformer networks have prompted a shift towards transformer-based or hybrid approaches, demonstrating enhanced performance and flexibility in various tasks.

In scenarios where privacy is a significant concern, such as applications involving sensitive data, using face images for model training poses challenges. Additionally, in domains like driver monitoring, the preference for eye images over face images is driven by the necessity for real-time estimation on edge devices with limited computational resources. These constraints underscore the need for robust and efficient gaze estimation methods that can operate under diverse conditions.

Motivated by the recent success of hybrid transformer-convolutional networks, we propose a novel dual encoder approach that departs from traditional methods. Unlike the approach in [5], which employs image enhancement as a pretraining step, our method integrates image enhancement as an auxiliary and parallel task. This strategy allows us to fully exploit the potential of image enhancement throughout the training process, enhancing the overall feature representation.

Moreover, we introduce an auxiliary network designed to predict task-specific uncertainty, which is then utilized to weight the losses for each task. This nuanced approach not only improves the accuracy of gaze estimation but also provides a more comprehensive framework for addressing the complexities inherent in this problem.

The core contributions of our work are as follows:

- **Dual Hybrid Encoder Architecture**: We propose a dual encoder framework combining Vision Transformers and Convolutional Neural Networks, leveraging their complementary strengths for enhanced gaze estimation.
- **Differentiable Gaussian Heatmap Loss**: We introduce a novel heatmap loss that spatially encodes gaze vectors, improving gaze localization accuracy.
- **Fourier Encoding-Based Loss:** We develop a fourier encoding-based loss to capture high-frequeny components, enhancing the sensitivity to fine gaze variations.
- **Uncertainty-Based Loss Weighting**: We incorporate an auxiliary network to predict task uncertainty, dynamically adjusting loss weights to improve model robustness and accuracy.
- **State-of-the-Art Performance**: Our method achieves state-of-the-art results on the MPIIGaze and RT-GENE datasets, demonstrating its effectiveness in diverse conditions and applications.

2 Related Work

Gaze estimation methods have traditionally relied on analyzing physical features such as ocular movement patterns, fixations, saccades [11], and smooth pursuits, due to their straightforward association with eye movement. As the understanding of eye geometry evolved, model-based approaches emerged, focusing on geometric features like the pupil center and eye contours [20]. These methods aimed to improve precision by leveraging detailed geometric characteristics. With the rise of deep learning, appearance-based methods using face or eye images for end-to-end gaze prediction gained popularity [26]. These can be categorized into face-appearance and eye-appearance-based methods.

Face Appearance-Based Gaze Estimation: Krafka et al. [16] pioneered gaze estimation using face images along with eye images, cropping regions of interest using facial detectors. Cheng et al. [8] assigned weights to eye features guided by facial features. Gaze-360 [13] introduced a pinball loss function to predict error quantiles, indicating prediction confidence. Abdelrahman et al. [1] converted the regression problem into a classification problem by dividing the gaze range into discrete bins. D'Antimo et al. [19] used an attention branch parallel to feature extraction for weighted eye features combined with face features. Transformer models [6, 30] have been explored for their efficacy in generating diverse features. GazeTR [5] proposed a hybrid transformer model incorporating convolutional layers to extract low-level features, passed onto transformer layers.

Recent efforts include using eye landmarks [29] and pupil centers [17] to enhance accuracy. FAZE [21] employed an autoencoder to learn compact representations of gaze, head pose, and appearance, introducing geometric constraints and meta-learning for adaptable gaze estimation. Facial appearance provides contextual benefits but is not always feasible for large networks, especially in AR/VR applications that require direct eye gaze estimation.

Eye Appearance-Based Gaze Estimation: Zheng et al. [31] initiated the use of single-eye grayscale images with head pose data for gaze direction estimation. Networks based on VGG-16 or ResNet have improved accuracy in this field [32]. Kang Il et al. [9] enhanced accuracy using features from both eyes. [7] introduced a four-stream network for feature extraction from both eyes. Chen et al. [4] used dilated convolutions for high-level feature extraction, increasing the receptive field without losing spatial resolution. Singh et al. [25] proposed a multitask architecture with image enhancement and adaptive binning for gaze regression. Recent attention has focused on gaze redirection [12, 24], synthesis, and scene-based gaze understanding.

This work focuses on gaze estimation from eye appearance, arguing that a hybrid architecture combining transformers and convolutional networks offers better representation. Using a combination of different losses to represent variations in gaze localization, we aim to achieve higher accuracy across diverse use-cases.

Supervision metrics: Predicting heatmaps for keypoint estimation has been explored [28], but for gaze estimation, direct correlation with pixel features is challenging. We argue that heatmap-based approaches provide initial localization by creating a heatmap from the gaze vector as a representative loss instead of predicting it directly. Using high-frequency features for better supervision, as seen in works utilizing Fourier transforms for super-resolution [10] or drawing parallels with Wasserstein loss for classification [2], leads us to propose a novel loss representation. By uplifting 2D gaze vectors to higher dimensions with Fourier encoding, we achieve a more granular metric. Inspired by uncertainty prediction tasks [14, 15], we propose an Uncertainty-Driven task importance weighting for dynamic loss adjustment.

3 Method

In this section, we detail the proposed architecture and loss functions designed to enhance eye gaze estimation accuracy and robustness. Our approach integrates a Vision Transformer and Convolutional Neural Network within a multi-branch framework, augmented with novel loss functions to leverage both spatial and frequency domain features effectively.

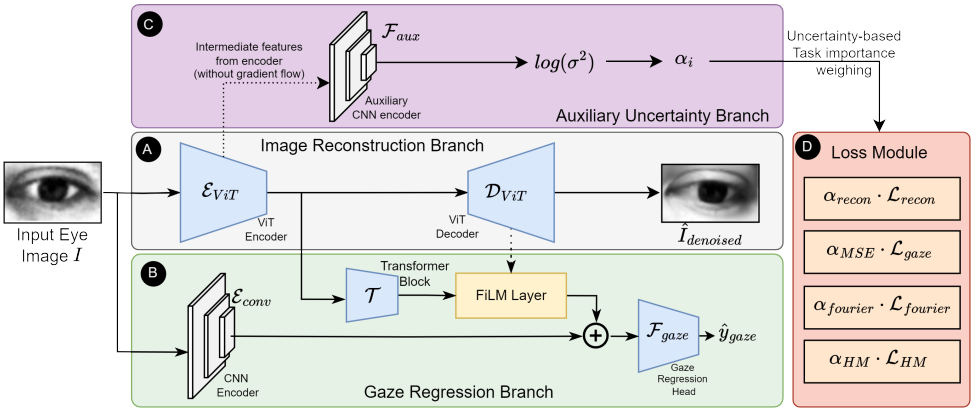

Figure 1: An outline of the proposed approach towards eye gaze estimation. The three main components comprise of (A) An image reconstruction and denoising branch, (B) The gaze estimation regression branch, and (C) An auxiliary uncertainty branch. Finally, in (D) the supervision comes from a set of losses weighed by the predicted uncertainty.

3.1 Architecture Overview

The proposed architecture for eye gaze estimation integrates three main branches—reconstruction, gaze estimation, and auxiliary uncertainty—to leverage the complementary strengths of Vision Transformers and Convolutional Neural Networks, enhancing both feature representation and prediction robustness. The overall approach is highlighted in Fig. 1.

A. Reconstruction Branch: is designed to leverage a Vision Transformer encoder $\mathcal{E}_{ViT}$ and decoder $\mathcal{D}_{ViT}$ for reconstructing the input eye image $I \in \mathbb{R}^{C\times W\times H}$. The output of this branch is the reconstructed image $\hat{I}$. This branch not only focuses on image reconstruction but is also configured to perform denoising tasks. This dual functionality ensures that the model learns robust and enhanced feature representations. Notably, this branch does not rely on pretraining for reconstruction but instead is utilised as a task alongside gaze regression, allowing the model to preserve structural information about the eye during overall training.

B. Gaze Estimation Branch: comprises a Convolutional Neural Network (CNN) encoder $\mathcal{E}_{conv}$ that processes the eye image I. Simultaneously, a transformer block $\mathcal{T}$ processes encoded image features $h^{\mathcal{E}_{ViT}} = \mathcal{E}_{ViT}(I)$ from the ViT encoder, resulting in $h^T = \mathcal{T}(h^{\mathcal{E}_{ViT}})$. We further enhance these features by applying Feature-wise Linear Modulation (FiLM) [23], which uses intermediate features $h^{\mathcal{D}_{ViT}}$ from the decoder to generate shift and scale parameters, yielding modulated features $h^F = FiLM(h^T, h^{\mathcal{D}_{ViT}})$. These FiLM-modulated features are concatenated with the features $h^{\mathcal{E}_{conv}}$ from the CNN encoder and passed through the gaze regression head $\mathcal{F}_{gaze}$. The final output is the predicted gaze vector $\hat{y}_{gaze} = \mathcal{F}_{gaze}(h^F; h^{\mathcal{E}_{conv}})$, which consists of the gaze yaw ($\hat{\theta}$) and pitch ($\hat{\phi}$). This hybrid approach exploits the strengths of both Vision Transformers, which capture dense spatial relations through self-attention, and CNNs, which are adept at learning low-level image features.

C. Auxiliary Uncertainty Branch: To enhance the robustness of our predictions, we introduce an auxiliary network $\mathcal{F}_{aux}$ that predicts uncertainty-based weights for the combined loss functions used in training. This auxiliary network utilizes re-aligned intermediate features from the encoder $\mathcal{E}_{ViT}$ to predict task-level log-variance $\log(\sigma_i^2)$. We adjust the architecture to ensure positivity of these values and convert them to loss weights $\alpha_i = \exp(-\log(\sigma_i^2))$, where i is the index of the loss metrics. The encoder features used by the auxiliary network are frozen during training to prevent direct gradient updates to the encoder parameters. This mechanism allows the auxiliary network to dynamically adjust the loss weights based on the image features, optimizing the balance between the gaze estimation and reconstruction tasks. Importantly, this auxiliary network is only active during training and is not required during the inference phase.

This architecture is designed to ensure that the model not only accurately reconstructs the input images but also precisely predicts gaze direction with a quantified measure of uncertainty, enhancing both the reliability and interpretability of the predictions.

3.2 Loss Metrics

In this subsection, we discuss the loss metrics utilized in training the proposed architecture and the impact of loss weighting through the auxiliary network.

Reconstruction Loss: Given the input image with applied augmentations and corruptions I, the predicted image $\hat{I}$, and the ground truth image I^*, the reconstruction loss is formulated as $\mathcal{L}_{recon} = ||I^* - \hat{I}||_2^2$ over the entire image. This reconstruction loss facilitates self-supervised training of the feature encoder to learn the image structure comprehensively. Through self-attention mechanisms, it provides rich features from the encoder that are subsequently utilized in the gaze estimation pipeline.

Gaze Heatmap Loss: In typical approaches [28], a heatmap is directly predicted by the network for keypoint estimation, however such an approach is not feasible for gaze estimation since the position of gaze projection does not correlate well with the pixel features of the input image. We employ a heatmap-based loss strategy to encode the predicted and true gaze vectors spatially as projected vectors on a plane in front of the eye origin, and compute the error between these projections. We utilize a differentiable Gaussian heatmap formulation to convert the 2D gaze vectors, projecting the yaw and pitch (θ, ϕ) onto a 2D heatmap $H \in \mathbb{R}^{M \times M}$ for both prediction and true gaze:

$$H(i,j) = \exp\left(-\frac{(i-\bar{\theta})^2 + (j-\bar{\phi})^2}{2\varepsilon^2}\right); \bar{\theta} = \frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}}, \bar{\phi} = \frac{\phi - \phi_{min}}{\phi_{max} - \phi_{min}} \tag{1}$$

Here, ε is the hyperparameter controlling the Gaussian spread, and $(\bar{\theta}, \bar{\phi})$ are the normalized heatmap coordinates of gaze projection based on gaze angular ranges. Adjusting the spread and scale hyperparameters during training allows the network to adaptively achieve higher precision as training progresses. The heatmap loss is thus defined as $\mathcal{L}_{HM} = ||H - \hat{H}||_2^2$.

Gaze Fourier Loss: In addition to the spatial mapping of the gaze vectors, we propose another representation of gaze prediction to facilitate fine-grained tuning of precision in the regression task using a Fourier encoding method. Fourier encoding of features allows for

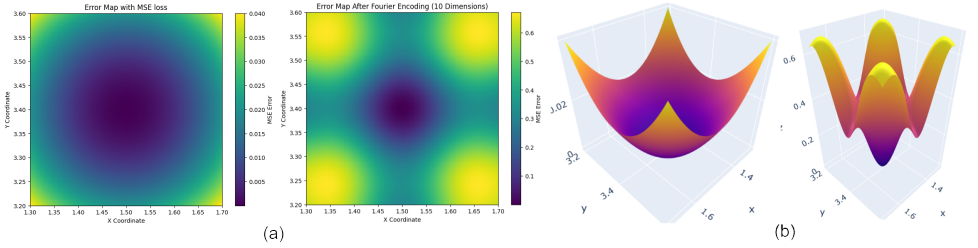

Figure 2: A visualization of the impact of fourier encoding on the L2 loss. We consider a 2D point (say ground truth) and the local region near the point, then visualize the loss in the local region (a, left) with, and (a, right) without the fourier loss combined with the L2 loss. It is clear that in the local region, the fourier loss is more susceptible to small changes and results in bigger gradient and faster convergence.

numerical stability and enhanced representation capability. The Fourier encoding is applied to the gaze vector as:

$$\hat{f}(\theta,\phi,B) = \{[\sin(2\pi b_i\theta),\cos(2\pi b_i\theta)] \mid b_i \in B\} \tag{2}$$

where B denotes the set of frequency bands chosen as a hyperparameter during training. The Fourier-integrated loss is then defined as:

$$\mathcal{L}_{fourier} = ||\hat{f}(\theta,\phi,B) - \hat{f}(\hat{\theta},\hat{\phi},B)||_2^2 \tag{3}$$

We argue that representing the gaze vectors in Fourier-encoded space enhances sensitivity to finer deviations between the prediction and true gaze. This is particularly beneficial in the later stages of training when predictions closely align with the ground truth, and the error scales reduce significantly, making the Fourier-based loss crucial for further optimization. A visual example of how the fourier loss affects the L2 loss when combined is also demonstrated in Fig. 2.

Gaze Regression Loss: is computed between the predicted gaze vector $\hat{y}_{gaze}$ and the ground truth gaze vector y_{gaze} as the L2 loss, similar to the reconstruction loss, $\mathcal{L}_{gaze} = ||y_{gaze} - \hat{y}_{gaze}||_2^2$. When combining the regression loss with the Fourier-based loss, we apply an additional weighting between the L2 regression and Fourier losses. The Fourier loss is sensitive to fine deviations in the prediction and may be detrimental during the initial training stages. Therefore, we define the weighting factor λ as inversely proportional to the gaze regression loss magnitude, $\lambda \propto |\mathcal{L}_{gaze}|^{-1}$ or $\lambda = k \cdot |\mathcal{L}_{gaze}|^{-1}$ where k is a scaling factor to keep $\lambda \in (0,1)$. The combined gaze and Fourier loss is then updated as:

$$\lambda \cdot \mathcal{L}_{fourier} + (1-\lambda) \cdot \mathcal{L}_{gaze} \tag{4}$$

This formulation allows higher weight to the L2 gaze loss in the initial training stages since $|\mathcal{L}_{gaze}|^{-1}$ is small and gradually shifts focus to the Fourier loss as the L2 gaze loss diminishes, indicating improved localization of the gaze vector.

3.3 Overall Loss

The overall loss is formulated as a weighted combination of the individual losses discussed above. For the proposed methods, we employ all the losses associated with reconstruction and gaze estimation, weighted by the updated predictions from the auxiliary network α_i. The comprehensive loss function is defined as follows:

$$\mathcal{L}_{total} = \sum_i \alpha_i \cdot \mathcal{L}_i + \log\left(\frac{1}{\alpha_i}\right) \tag{5}$$

This additional logarithmic term serves as a regularization component to stabilize the predicted loss weights, by preventing extreme weight values, aligning our formulation with the principles of heteroscedastic uncertainty as proposed in [15]. However, our approach diverges by applying the predicted weights and uncertainty estimates at the task level for different types of gaze estimation losses. This methodology not only enhances the robustness of the training process but also provides valuable insights into the model's training dynamics and progression. We further discuss the impact of these weights on task-level losses in the supplementary material.

4 Experiments & Results

4.1 Dataset Description

To rigorously evaluate our model, we utilized two prominent datasets known for their diversity in gaze direction, head pose, and appearance: MPIIGaze [32] and RT-Gene [9]. These datasets are critical for developing appearance-based gaze estimation methods.

MPIIGaze Dataset: is a comprehensive dataset provided by the Max Planck Institute for Informatics, comprising 213,659 images from 15 participants. The dataset captures variations in lighting conditions, head poses, and gaze targets, using everyday laptop cameras, thus making it highly applicable for human-computer interaction tasks. Each entry includes normalized eye images associated with corresponding face images, enhancing the dataset's utility for robust gaze estimation models. For evaluation purposes, MPIIGaze follows a standard protocol involving 3,000 images per subject, utilizing a leave-one-out cross-validation framework to assess model performance across different subjects.

RT-GENE Dataset: developed by the Technical University of Munich, focuses on real-time gaze estimation in outdoor settings. This dataset comprises 122,531 images from 17 subjects, captured under natural lighting with diverse backgrounds. It features high-resolution face images with annotated gaze directions and head poses, collected using gaze-tracking glasses. The dataset includes both inpainted and original images, totaling 229,116 distinct eye crops. Due to the reported noise in inpainted images [19], we exclusively use the original images for training and evaluation. RT-GENE employs a 3-fold cross-validation methodology for testing, ensuring comprehensive assessment across varied data splits.

By leveraging these datasets, our goal is to validate the robustness and accuracy of our proposed model under diverse and challenging real-world conditions, thereby enhancing its applicability in practical scenarios.

Datasets	ARENet	RT-GENE	AGENet	EG-SIF	Base ViT	Base ViT-Conv	Loss HM	Loss Fourier	Loss Combined	Auxiliary Weights
MPIIGaze	5.02	4.8	4.64	4.53	5.91	5.48	4.87	4.92	5.01	**4.43**
RT-GENE	-	8.6	7.44	7.41	9.41	8.63	7.57	7.48	7.61	**7.19**

Table 1: Angular error results on MPIIGaze and RT-GENE datasets. The *Auxiliary Weights* method represents the complete pipeline proposed in this work and shows superior performance.

4.2 Experiment Setup

We employ normalized eye images with a resolution of 36x60 pixels for training on both the MPIIGaze and RT-GENE datasets. For the MPIIGaze dataset, we enhance training efficiency by flipping the right eye images to geometrically align them with the left eye images, facilitating the model's learning of consistent features. Each eye image is used to predict an individual gaze value. In contrast, for the RT-GENE dataset, we concatenate the left and right eye images side by side, and the model is trained to predict the average gaze value for both eyes. The primary evaluation metric is the gaze angular error $\theta_{angular}$, consistent with standard practices in gaze estimation research. The training parameters are set with a learning rate of 1e-3, using a multistep learning rate scheduler, over 50 epochs, utilizing the Adam optimizer. We apply various augmentations, including ColorJitter, GaussianBlur, and normalization, to enhance the robustness and generalizability of the model.

To elucidate the various facets of the proposed pipeline and the impact of each component, we structure our approach into multiple progressive experiments, reporting the results for each set incrementally. This methodology allows us to isolate the individual contributions of each proposed component and underscore the relative improvements introduced by their integration.

The core components of the proposed pipeline include the Vision Transformer architecture and the convolutional encoder, working in tandem on the architectural side. On the loss method side, we incorporate the proposed heatmap-based loss, Fourier encoding loss, and the uncertainty-based weighting for loss combination. To ensure fair comparisons, we delineate the following experiment sets:

Base ViT experiment utilizes the base loss functions (reconstruction and L2 gaze) with only the Vision Transformer encoder, establishing a baseline. **Base ViT-Conv** experiment integrates the convolutional encoder, showcasing the relative enhancement achieved by adding convolutional features to the transformer-only architecture. **Loss HM** further incorporates the heatmap loss into the experiment pipeline, while **Loss Fourier** substitutes the heatmap loss with the Fourier encoded loss, highlighting the distinct impact of each loss function. **Loss Combined** aggregates all the proposed loss metrics, summing them directly without loss weighting. Finally, **Auxiliary Weights** represents the complete pipeline proposed in this work, incorporating dynamic loss weighting through the auxiliary network.

4.3 Result Analysis

The results are summarized in Table 1, showcasing the average angular error for each experimental setup on the MPIIGaze and RT-GENE datasets. Table 2 provides a detailed breakdown of person-specific angular errors and the average error per individual in the leave-one-out validation setting.

Experiments	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Average
Loss HM	3.4	3.68	3.44	5.42	4.94	5.23	4.74	6.23	5.48	5.15	5.9	3.85	4.86	5.2	5.54	4.87
Loss Fourier	3.54	4.16	4.38	5.41	5.05	5.22	3.86	5.94	5.47	5.07	5.56	4.16	4.81	5.41	5.79	4.92
Loss Combined	3.31	4.04	4.04	5.5	5.28	5.2	4.33	5.47	5.5	6.16	6.17	4.33	4.7	5.5	5.48	5.01
Auxiliary Weights	3.31	3.68	2.72	5.1	4.4	4.99	3.86	5.23	5.45	4.86	4.26	3.85	4.7	5.01	5.15	4.43

Table 2: Experiment results on each person in the MPIIGaze dataset with different experiment variations to understand the impact of our proposed methods.

Our comprehensive evaluation demonstrates that the complete proposed approach, as highlighted in the **Auxiliary Weights**, achieves the best results, establishing state-of-the-art performance on both the MPIIGaze and RT-GENE datasets.

The integration of the convolutional encoder notably enhances overall performance compared to a Vision Transformer-only approach, underscoring the benefits of combining convolutional features with transformer-based representations. This improvement is consistently observed across both datasets, validating the effectiveness of our hybrid architecture.

Furthermore, the introduction of dynamic loss weighting significantly contributes to performance enhancement. This component optimizes the balance between various loss terms, leading to more accurate gaze estimation. Our results also indicate substantial individual improvements with both the heatmap loss and Fourier encoding loss, supporting our hypothesis that these loss functions effectively enhance the model's sensitivity to gaze localization.

In Table 2, we again notice a similar pattern as shown in the average case, however slight deviations are present which can be attributed to the large differences in the type of image distributions available for each individual subject. We further explore the results for individual subjects with visualizations in the supplementary material.

In summary, the results confirm that each component of our proposed method—whether architectural enhancements or novel loss functions—contributes to the superior performance observed. The dynamic loss weighting, in particular, plays a crucial role in fine-tuning the model's focus, leading to state-of-the-art accuracy in gaze estimation tasks.

5 Conclusion

In this paper, we introduced a novel hybrid architecture for eye gaze estimation that synergizes Vision Transformers and Convolutional Neural Networks. By incorporating novel loss functions, including a differentiable Gaussian heatmap loss and a Fourier encoding-based loss, along with auxiliary uncertainty-based loss weighting, our approach demonstrates significant improvements in accuracy and robustness. Comprehensive evaluations on the MPIIGaze and RT-GENE datasets highlight the effectiveness of our method, establishing it as a state-of-the-art solution for gaze estimation. This work not only advances the precision of gaze prediction but also enhances the model's applicability in real-world scenarios, particularly in human-computer interaction and augmented reality contexts.

Acknowledgement

We wish to thank the organization Mercedes-Benz Research & Development India for providing resources for experiments such as GPU compute clusters.

References

[1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments, 2022.

[2] Gennaro Auricchio, Andrea Codegoni, Stefano Gualandi, and Lorenzo Zambon. The fourier loss function. *arXiv preprint arXiv:2102.02979*, 2021.

[3] David Beymer and Daniel M. Russell. Webgazeanalyzer: A system for capturing and analyzing web reading behavior using eye gaze. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, page 1913–1916, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930027. doi: 10.1145/1056808.1057055. URL https://doi.org/10.1145/1056808.1057055.

[4] Zhaokang Chen and Bertram E. Shi. Appearance-based gaze estimation using dilated-convolutions, 2019.

[5] Yihua Cheng and Feng Lu. Gaze estimation using transformer. *CoRR*, abs/2105.14424, 2021.

[6] Yihua Cheng and Feng Lu. Gaze estimation using transformer, 2021.

[7] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[8] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation, 2020.

[9] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 339–357, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01249-6.

[10] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution, 2021.

[11] Emile Javal. Essai sur la physiologie de la lecture. *Annales d'Ocilistique*, 80:97–117, 1878.

[12] Shiwei Jin, Zhen Wang, Lei Wang, Ning Bi, and Truong Nguyen. Redirtrans: Latent-to-latent translation for gaze and head redirection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5547–5556, 2023. URL https://api.semanticscholar.org/CorpusID:258822908.

[13] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild, 2019.

[14] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *CoRR*, abs/1703.04977, 2017.

[15] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115, 2017.

[16] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone, 2016.

[17] Kang Il Lee, Jung Ho Jeon, and Byung Cheol Song. Deep learning-based pupil center detection for fast and accurate eye tracking system. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 36–52, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58529-7.

[18] Peng Li, Xuebin Hou, Xingguang Duan, Hiuman Yip, Guoli Song, and Yunhui Liu. Appearance-Based Gaze Estimator for Natural Interaction Control of Surgical Robots. *IEEE Access*, 7:25095–25110, January 2019. doi: 10.1109/ACCESS.2019.2900424.

[19] LRD Murthy and Pradipta Biswas. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3143–3152, June 2021.

[20] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research &ampsemicolon Applications*. ACM, jun 2018. doi: 10.1145/3204493.3204545. URL https://doi.org/10.1145%2F3204493.3204545.

[21] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation, 2019.

[22] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David P. Luebke, and Aaron E. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35:1 – 12, 2016. URL https://api.semanticscholar.org/CorpusID:17614452.

[23] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[24] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields, 2023.

[25] Vasudev Singh, Chaitanya Langde, Sourav Lakotia, Vignesh Kannan, and Shuaib Ahmed. EG-SIF: Improving appearance based gaze estimation using self improving features. In *NeuRIPS 2023 Workshop on Gaze Meets ML*, 2023. URL https://openreview.net/forum?id=tY06Zwn3v4.

[26] Kar-Han Tan, D.J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002. (WACV 2002). Proceedings.*, pages 191–195, 2002. doi: 10.1109/ACV.2002.1182180.

[27] Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bertram E. Shi. Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1476–1479, 2015. doi: 10.1109/EMBC.2015.7318649.

[28] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019.

[29] Zhengyang Wu, Srivignesh Rajendran, Tarrence van As, Joelle Zimmermann, Vijay Badrinarayanan, and Andrew Rabinovich. Eyenet: A multi-task network for off-axis eye gaze estimation and user understanding, 2019.

[30] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan Yuille, and Wei Shen. Glance-and-gaze vision transformer, 2021.

[31] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2015. doi: 10.1109/CVPR.2015.7299081.

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation, 2017.

[33] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, may 2019. doi: 10.1145/3290605.3300646. URL https://doi.org/10.1145%2F3290605.3300646.