Supplementary – 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

We discuss the additional details about the proposed method and further analysis on the resutls in this supplementary material. We start with an understanding of the heatmap preparation for the localised heatmap-loss in Sec. 1, followed by further explaination of the fourier loss in Sec. 2. We then discuss about the impact of the Uncertainty-based auxiliary loss weighing strategy in Sec. 3. Finally, we discuss the results and subject-level analysis for MPIIGaze dataset in Sec. 4 and for RT-GENE dataset in 5.

1 Heatmap Loss Visualized

We utilise a heatmap based loss strategy in the loss formulation for additional supervision along with the L2 loss. For the heatmap, we prepare a gaussian function centered around the normalised gaze coordinates with a spread controlled by a hyper-parameter. This gaussian function is scaled in a way that the maximum value lies at the pixel corresponding to the gaze center and decreases following the gaussian function. The formulation has been discussed in the main paper as follows:

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

Where, $H(i,j)$ is the pixel coordinate on the heatmap,($\bar{\theta},\bar{\phi}$ are the normalised gaze vectors, ε is the hyper-parameter which controls the spread of the gaussian. The entire function is differentiable without any trainable parameters and can be used directly for training. We argue the usefulness of the heatmap strategy especially during the early phases of the training where the network first learns to predict the gaze in a localised area near to the ground truth but with lower accuracy.

A sample visualization for the eye-image, predicted and ground truth heatmap is shown in fig. 1. We notice that even when the gaze prediction may be slightly incorrect, the overlap between the gaussians results in a lower loss value overall and guides the prediction to a localised region near the true gaze. To the best of our knowledge, such use of heatmap for

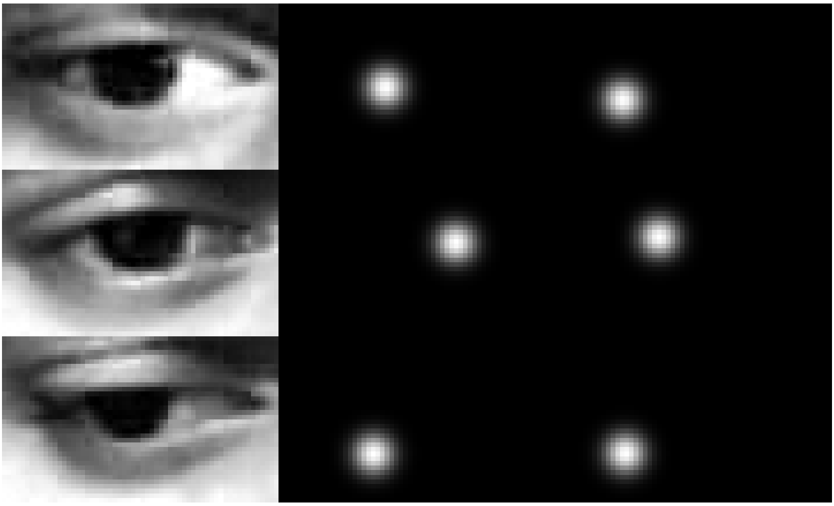

Figure 1: The columns represent the eye image, heatmap for the ground truth, and the predictions respectively. The center of each gaussian disc represents the normalised location of the gaze vector projected onto a plane in front of the eye.

loss formulation rather than a prediction task directly is novel and helps the training process. Furthermore, as is clear from the visualizations in fig. 1, the location of the heatmap does not share a pixel-level visual correspondance with the eye image directly since the heatmap represents the gaze vector projected on to a plane in front of the eye. This makes direct heatmap regression ill-posed from eye images. We further discuss the role of the heatmap loss in the results section.

2 Fourier Loss Analysis

The Fourier encoding method transforms gaze vectors into a higher-dimensional space, capturing fine-grained variations in the data. Given a gaze vector characterized by yaw (θ) and pitch (ϕ), we apply Fourier encoding as follows:

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

where B represents a set of frequency bands selected as hyperparameters during training. This transformation ensures numerical stability and enhances representation capability, particularly useful for capturing periodic and high-frequency components of gaze vectors.

Fourier encoding leverages the principles of signal processing to capture periodic functions, making it particularly suitable for gaze estimation tasks. By representing gaze vectors in a higher-dimensional space, Fourier encoding enhances the model's ability to learn complex patterns. Our ablation studies reveal the significant impact of Fourier loss on model performance. Specifically, models incorporating Fourier loss demonstrate improved convergence rates and enhanced precision in gaze estimation tasks. These improvements are particularly evident in the later stages of training, where finer adjustments are critical.

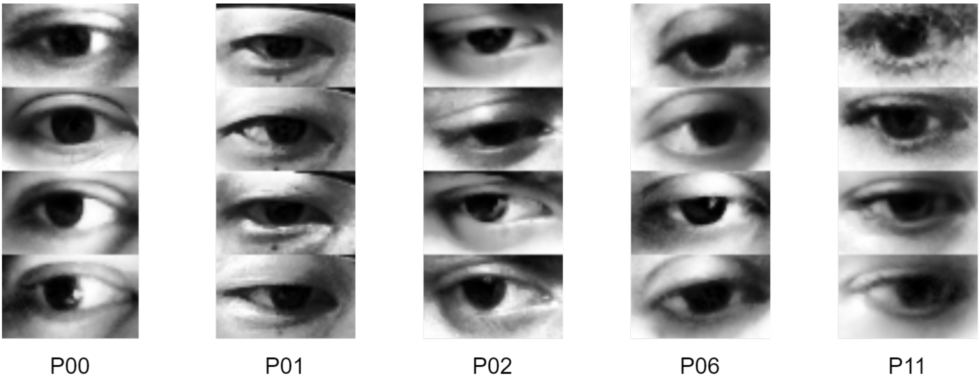

Figure 2: Some examples of the clear images in the MPIIGaze dataset. The eye images are clear and sharp, providing descriptive features for the prediction task.

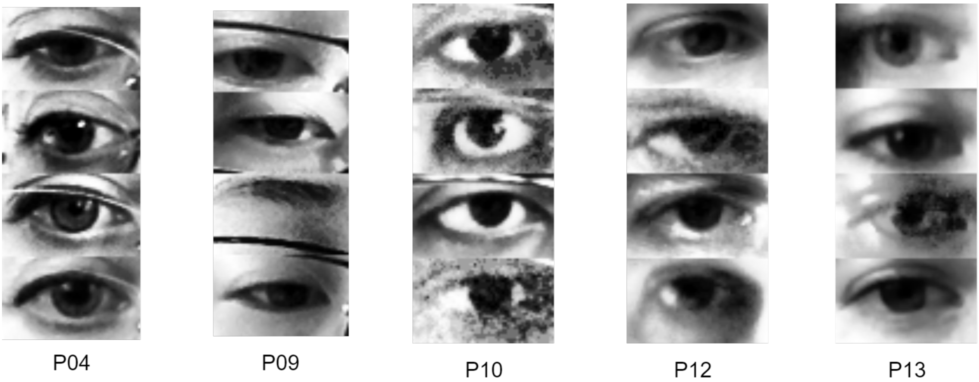

Figure 3: Some examples of the slightly noisy images in the MPIIGaze dataset. These images contain some degree of noise in the images or may slight occlusion due to the presence of eye glasses and frames. Such artifacts may pose some challenge towards the predictive task, however the denoising reconstruction network is able to extract robust features from such images.

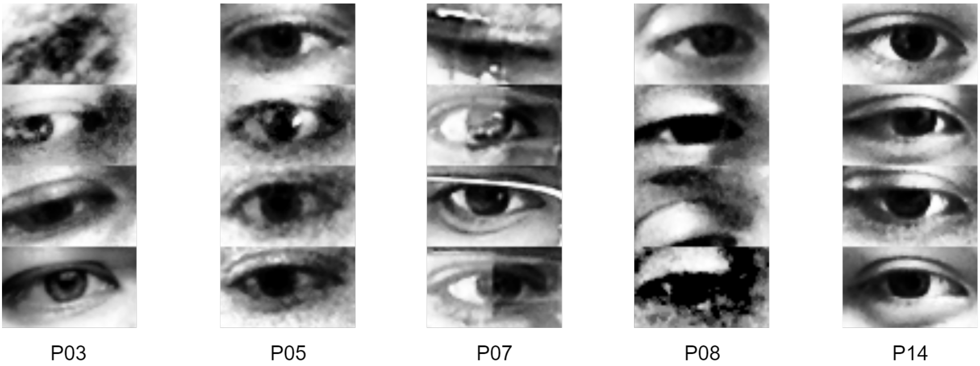

Figure 4: Examples of the noisy/occluded/corrupted images in the MPIIGaze dataset. We especially notice glasses and frames present in the images along with reflections on the surface which pose significant challenge. Additionally, noticable alignment and offset challenges make some of these images especially difficult for the model to process.

3 Auxiliary Loss Weighing

As mentioned in the main paper, the overall loss is formulated as a weighted loss using the weights predicted by the auxiliary network (α_i):

$$\mathcal{L}_{total} = \sum_i \alpha_i \cdot \mathcal{L}_i + \log\left(\frac{1}{\alpha_i}\right) \tag{3}$$

The auxiliary network predicts task-specific uncertainty weights that dynamically adjust the importance of each loss component during training. This dynamic adjustment ensures that the model focuses on the most relevant aspects of the task at different stages of training, improving overall performance and robustness.

The predicted weights from the auxiliary network play a crucial role in shifting the focus between the heatmap and Fourier losses, especially during the initial and final phases of training. For example, in the early stages, the model benefits from focusing on good localization of the gaze prediction within an approximate area. This is reflected by an increase in the weights assigned to the heatmap loss. As training progresses and the model's predictions become more refined, the focus shifts towards capturing finer details and high-frequency components of the gaze vectors, leading to an increase in the weights assigned to the Fourier loss. This pattern in the change of auxiliary weights is highlighted in fig. 6.

4 Results: MPIIGaze

MPIIGaze Dataset: from the Max Planck Institute for Informatics consists of 213,659 images from 15 people. It captures lighting conditions, head poses, and gaze targets via laptop cameras, making it useful for human-computer interaction tasks. Entries include normalized eye and face images, aiding in robust gaze estimation models. Evaluation involves 3,000 images per subject in a leave-one-out cross-validation framework.

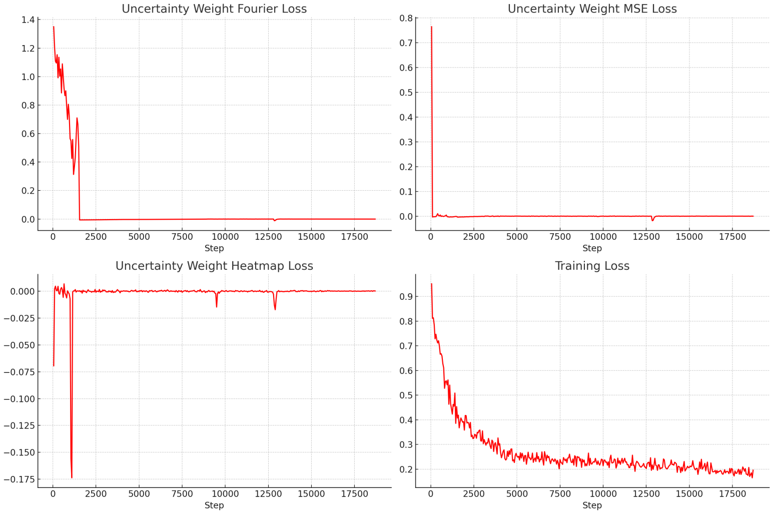

Figure 5: The predicted loss weighings visualized against the training steps. The weights for a) Fourier loss, b) L2 regression loss, c) Heatmap loss, and d) Overall loss express the shift of focus during the training as discussed in sec 3. The magnitude of heatmap weightage is higher during the initial phase of training and decreases as training progresses towards convergence.

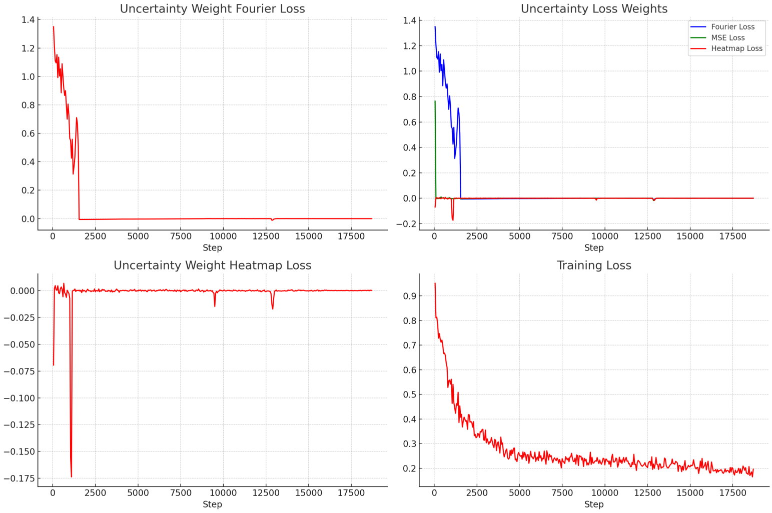

Figure 6: The predicted loss weighings visualized against the training steps for another training example. The weights for a) Fourier loss, b) Loss Weights, c) Heatmap loss, and d) Overall loss express the shift of focus during the training as discussed in sec 3.

Person	Base ViT	Base ViT-Conv	Loss HM	Loss Fourier	Loss Combined	Auxiliary Weights	HParams Tuning
00	4.38	3.84	3.4	3.54	3.31	3.31	**3.27**
01	6.63	5.67	3.68	4.16	4.04	3.68	**3.06**
02	4.26	4.23	3.44	4.38	4.04	**2.72**	**2.72**
03	6.3	5.83	5.42	5.41	5.5	5.1	**5.09**
04	5.7	5.74	4.94	5.05	5.28	4.4	**4.02**
05	5.82	5.58	5.23	5.22	5.2	**4.99**	**4.99**
06	6.03	5.83	4.74	**3.86**	4.33	**3.86**	**3.86**
07	7.12	6.71	6.23	5.94	5.47	5.23	**5.12**
08	6.2	5.64	5.48	5.47	5.5	5.45	**5.03**
09	6.8	5.46	5.15	5.07	6.16	4.86	**4.67**
10	5.71	5.3	5.9	5.56	6.17	4.26	**4.26**
11	5.47	5.12	**3.85**	4.16	4.33	**3.85**	**3.85**
12	6.23	5.9	4.86	4.81	4.7	4.7	**4.47**
13	6.24	5.77	5.2	5.41	5.5	5.01	**4.74**
14	5.73	5.53	5.54	5.79	5.48	5.15	**5**
Avg	5.91	5.48	4.87	4.92	5.00	4.43	**4.28**

Table 1: Degree errors for MPIIGaze dataset for each person in the cross-person training-validation setting. We show best overall results with the auxiliary weights + hyper-parameter tuning in the proposed approach.

Person	lr	batch	emb-dim	n_band	max_freq	patch
00	3e-4	32	64	8	10	3
01	9e-4	32	32	2	10	4
02	3e-4	128	32	6	6	6
03	3e-4	128	32	6	4	6
04	1e-3	256	128	8	6	6
05	1e-4	128	32	6	6	6
06	3e-4	256	64	8	6	4
07	1e-3	256	16	6	10	6
08	3e-4	128	16	10	2	6
09	1e-3	64	32	4	2	4
10	3e-4	128	32	6	6	6
11	3e-4	64	32	6	2	6
12	8e-4	64	32	8	4	12
13	1e-4	128	128	6	6	2
14	7e-3	256	16	6	8	3

Table 2: Hyper-parameter set for each person corresponding to the best score achieved in the HParams training experiment.

Hyperparameter Tuning: optimizes the model performance by exploring a search space to minimize the loss and boost accuracy. Adjusting parameters like batch size, learning rate, etc. helps in identifying the best model configuration. Effective tuning improves model robustness across data conditions, achieving a low average degree error and high performance across different image quality. We choose the following hyperparameters for optimizing the model configuration:

- Batch Size: This refers to the size of the mini-batch used for weight updates. The search space included $[32, 64, 128, 256, 512]$.
- Learning Rate: This controls the step size during the optimization process. The search space for the learning rate was $(1e-5, 1e-2)$.
- Embedding Dimension of the Transformer: This defines the size of the vector space in which the input is embedded. The search space was $[16, 32, 64, 128]$.
- Patch Size for Transformer: This parameter determines the size of patches extracted from the input image. The search space included $[2, 3, 4, 6, 12]$.
- Max Frequencies in Fourier Loss: This parameter influences the range of frequencies considered in the Fourier loss function. The search space was $[2, 4, 6, 8, 10]$.
- Number of Bands in Fourier Loss: This specifies the number of frequency bands used in the Fourier loss. The search space was $[2, 4, 6, 8, 10]$.

The results for MPIIGaze after hyperparameter tuning is summarized in Table 1 and the hyperparameter are listed in table 2. Our method achieves a state-of-the-art average degree error of **4.28**.

Result Analysis: As can be seen from 1, the degree error across different subjects can be divided into 3 brackets, $[(<4^\circ), (4^\circ - 5^\circ), (>5^\circ)]$. On further investigation these brackets can be linked to the image quality of the eye crops, as shown in Figures 2, 3, 4.

- Set-1 $(err < 4^\circ)$: comprises of subjects 00, 01, 02, 04, 06, as shown in Figure 2. The quality of eyes crops here is fairly good without much noise.
- Set-2 $(4^\circ < err < 5^\circ)$: comprises of subjects 04, 09, 10, 11, 12, as shown in Figure 3. The eye crops contains some amount of noise due to glasses, illumination, shawdows, etc.
- Set-3 $(err > 5^\circ)$: comprises of subjects 03, 05, 17, 08, 14, as shown in Figure 4. The eye crops here are highly noisy and hence makes it difficult to accurately estimate gaze.

The categorization based on image quality brackets allows for a nuanced understanding of how different image qualities impact the degree error. This stratification provides deeper insights into the performance dynamics of our model under varying conditions.

In summary, the results underscore the importance of meticulous hyperparameter tuning and also help us in understanding of dataset based on image quality. Our method not only achieves a low average degree error but also demonstrates robustness across different image quality brackets, confirming the effectiveness of our approach in handling diverse data conditions.

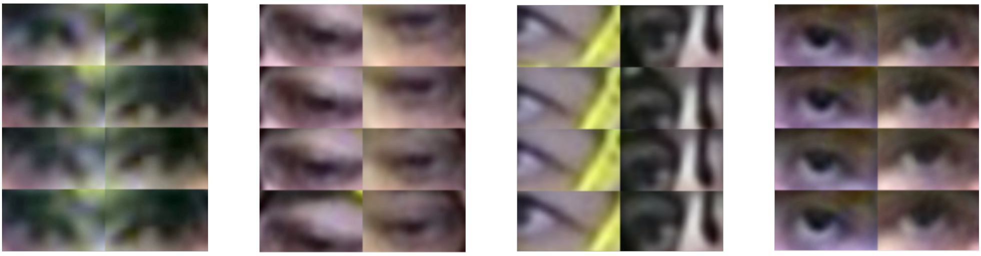

Figure 7: Eye image samples from the RT-GENE dataset. While the RT-GENE dataset provides RGB images, noticeable degradations are available in the dataset especially in the quality of the images and the position of the eye center.

5 Results: RT-GENE

RT-GENE: from the Technical University of Munich focuses on real-time outdoor gaze estimation. It includes 122,531 images from 17 subjects, all captured under natural lighting. The dataset contains high-resolution face images along with eye crops annotated with gaze directions and head poses, gathered using gaze-tracking glasses. We only use the original eye crops due to noise in the inpainted ones. RT-GENE uses a 3-fold cross-validation for testing. $S-1$ consists of subjects: ['5', '6', '11', '12', '13'], $S-2$ consists of subjects: ['3', '4', '7', '9'] and $S-3$ consists of subjects: ['1', '2', '8', '10'].

Test Sets	Base ViT	Base ViT-Conv	Loss HM	Loss Fourier	Loss Combined	Auxiliary Weights	HParams Tuning
S-1	8.34	6.92	7	6.85	7.12	**6.74**	**6.74**
S-2	9.76	8.85	7.81	7.74	7.72	7.35	**6.98**
S-3	10.13	8.62	8.02	7.86	8	7.5	**7.00**
Avg	9.41	8.13	7.61	7.48	7.61	7.19	**6.91**

Table 3: Degree errors on the RT-GENE dataset for each evaluation set in the dataset. We report the state-of-the-art results with the final hyper-parameter tuning experiment.

Test Set	lr	batch	emb-dim	n_band	max_freq	patch
S-1	1e-3	128	32	6	6	6
S-2	1e-3	64	16	8	10	4
S-3	5e-4	64	16	8	8	6

Table 4: Hyper-parameter set for evaluation set corresponding to the best score achieved in the HParams training experiment.

The results for RT-GENE after hyperparameter tuning is summarized in Table 3 and the hyperparameter are listed in table 4. We use a same hyperparameter space that we used for MPIIGaze [Sec 4]. Our method achieves a state-of-the-art average degree error of **6.91**.

As can be seen from Fig 7 that the eye crops are highly noisy as compared to those in MPIIGaze 4, due to higher variation in distance, occlusion due to tracking glasses, etc. and hence the higher average degree error as compared to MPIIGaze.

References