GazeOnce360: Fisheye-Based 360° Multi-Person Gaze Estimation with Global-Local Feature Fusion

Supplementary Material

1. MPSGaze360 Dataset Details

This section provides additional information about the MPSGaze360 dataset. We supplement the main paper with the following details.

1.1. MetaHuman Model Diversity

The dataset includes 69 MetaHuman digital models. Statistics of their gender, age, skin tone, and ethnicity distributions are summarized in Table 1. All attribute annotations were automatically inferred by the GPT-4o model using front-view facial images. While factors such as illumination variation, visual ambiguity, and hallucination may introduce deviations from expert-defined labels, the aggregated statistics offer a reasonably faithful characterization of the overall distribution based on our observations.

Table 1. Demographic attribute statistics of the 69 MetaHuman characters used in the MPSGaze360 dataset.

Attribute	Category	Count	Percentage
Gender	Female	32	46.38%
	Male	37	53.62%
Age	15–24 years	10	14.49%
	25–44 years	44	63.77%
	45–64 years	8	11.59%
	65 years and above	7	10.14%
Skin Tone	Fair	9	13.04%
	Light Medium	26	37.68%
	Medium	19	27.54%
	Deep Medium	12	17.39%
	Dark	3	4.348%
Ethnicity	African	23	33.33%
	Caucasian	21	30.43%
	East Asian	16	23.19%
	Other	9	13.04%

1.2. Diversity of Rendered Facial Appearance

Figure 2 presents close-up crops of faces from the dataset. The rendered images contain clear facial and ocular details, which are beneficial for learning causal gaze-related features. Across samples, we observe substantial diversity in lighting conditions, identity appearance, head pose, gaze direction, subject–camera distance, and facial sharpness.

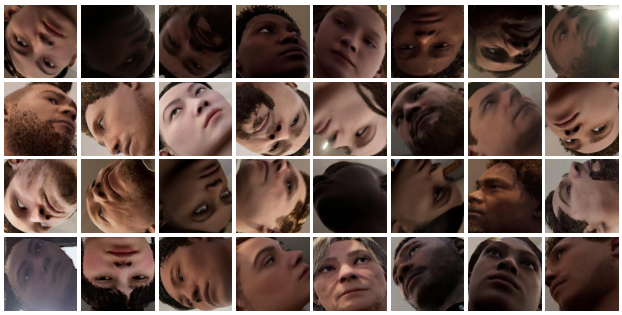

Figure 1. Examples of close-up facial crops from the MPSGaze360 dataset, illustrating the diversity in appearance, lighting, head pose, and gaze direction.

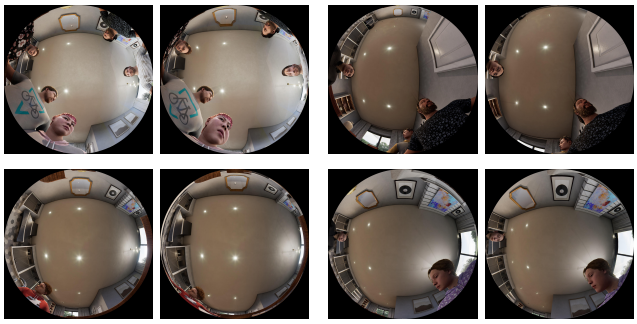

Figure 2. Comparison between images rendered using an equidistant projection and those rendered with a Kannala–Brandt (KB) distortion model.

1.3. Extended Fisheye Distortion Parameters

Figure 2 further compares images rendered using an equidistant projection with those produced using a Kannala–Brandt distortion model with a larger distortion coefficient. In practice, many commercially available fisheye cameras exhibit relatively mild distortion, so models trained on equidistant fisheye images over a cropped 180° FOV are often sufficient for gaze estimation. However, for cameras with stronger distortion, our synthesis pipeline allows distortion parameters to be adjusted directly based on camera calibration, enabling the generation of new images and corresponding ground-truth annotations. Importantly, this process does not require regenerating the entire scene; a fast transformation can be applied using the stored 3D ground-truth. This flexibility allows the dataset to support rapid adaptation to diverse fisheye cameras in deployments.

Table 2. Impact of different hyperparameters and loss weights.

Setting	Precision ↑	Recall ↑	Gaze Error (°) ↓	Adjusted Gaze Error (°) ↓
Baseline (lr=10^{-3}, glw=1.0)	0.9992	0.9903	10.39	10.20
lr=5×10^{-4}	0.9975	0.9905	11.85	11.74
lr=2×10^{-3}	0.9781	0.9574	13.16	13.14
glw=0.5	0.9877	0.9911	11.15	10.97
glw=5.0	0.9928	0.9899	8.742	8.503
glw=10.0	0.9989	0.9911	7.741	7.592

2. More Experiments

2.0.1. Robustness Across Face Resolutions

To evaluate robustness under varying face scales, we divide test samples into groups according to the detected face width and compute gaze errors for each group. As shown in Table 3, low-resolution faces (30–60 px) lead to larger angular errors due to reduced detail.

Table 3. Gaze error (°) under different face width intervals (pixels).

Setting	**30–60**	**60–90**	**90–120**	**120–150**	**>150**
w/o RotConv, w/o ldmks	13.91	11.65	12.02	11.75	11.70
RotConv only	12.50	10.73	11.09	10.66	12.17
RotConv + face ldmks	11.27	9.230	9.957	9.323	10.59
RotConv + eye ldmks	11.03	8.293	8.617	9.368	7.569
RotConv + face + eye ldmks	10.80	8.431	8.811	8.919	8.052

2.0.2. Metrics w.r.t Head Pose and Distance

Tables 4 and 5 report the gaze estimation errors under different head pose yaw angles and head distances, respectively. Specifically, Table 4 shows the gaze error across absolute yaw intervals from 0° to 90°, while Table 5 presents errors for varying head distances in centimeters.

The results indicate that gaze error increases moderately with larger head rotations and greater distances, reflecting the inherent difficulty of predicting gaze under extreme poses or when faces occupy fewer pixels.

Table 4. Gaze error (°) under different absolute head yaw angles.

Head pose yaw	**0–10**	**10–20**	**20–30**	**30–45**	**45–60**	**60–90**
	9.99	10.23	10.68	10.13	10.54	11.07

Table 5. Gaze error (°) under different head distances (cm).

Head distance	**30–50**	**50–70**	**70–90**	**>90**
	9.506	9.875	10.47	12.73

2.1. Hyperparameters and Loss Weights

In the main experiments, we use an initial learning rate of 10^{-3} and set all loss weights to 1 to avoid any task-specific weighting heuristics. In the supplementary material, we additionally retrain the model with different learning rates and gaze loss weights (glw), with results reported in Table 2. We observe that adjusting the learning rate generally has a negative impact on performance, whereas increasing the gaze loss weight can further improve the model's accuracy.

3. Additional Discussions

The additional experiments presented in this supplementary material further validate the robustness and generalization of GazeOnce360 under various settings, including different hyperparameters, loss weights, head poses, and distances. These results provide a deeper understanding of the model's behavior and sensitivity, complementing the main paper.

We also observe that the fisheye image resolution limits its gaze estimation for subjects located more than 2 meters from the camera, which in turn constrains the practical deployment of the system. Moreover, accurate estimation requires that subjects face the camera to some extent; extreme gaze angles relative to the camera, as well as occlusions between individuals, can lead to unreliable or missing gaze predictions. These factors represent inherent limitations of the current system.

Future work could explore real-world deployment considerations, domain adaptation techniques, fisheye gaze target estimation, and ethical safeguards.

4. Broader Impact

From a broader impact perspective, 360° multi-person gaze estimation has potential applications in human-computer interaction, social behavior analysis, and collaborative robotics. At the same time, as with all vision-based human analysis systems, care must be taken to respect privacy and prevent misuse in sensitive contexts.