GazeREC-Net: Advancing Gaze Restoration in Low-Light Conditions

Jiayin Ku, Li Wang

***Abstract*—Gaze estimation technology is essential for applications such as human-computer interaction, augmented reality, and virtual reality. However, its accuracy is significantly compromised in low-light conditions due to degraded image quality. To address this, we developed GazeREC-Net, an innovative gaze restoration method. We simulated low-light conditions on the MPIIFaceGaze and ColumbiaGaze datasets, creating a specialized degraded dataset for training and testing our model. GazeREC-Net combines Fourier transform techniques with advanced image restoration algorithms, significantly enhancing image quality and optimizing gaze information recovery and extraction in low-light environments. Our evaluations using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics demonstrated a PSNR improvement of 7.64% and an SSIM improvement of 5.75% compared to HINet. Additionally, GazeREC-Net outperformed CFTNet and other existing gaze estimation models, including Gaze-TR, Dilated-Net, CA-Net, L2CS-Net, and MTGLS, in reducing gaze estimation error. These findings validate the effectiveness of GazeREC-Net in low-light conditions and offer new research directions for applying gaze estimation technologies in complex lighting environments.**

***Index Terms*—Gaze Estimation, Low-Light Conditions, Image Restoration, Fourier Transform, Gaze Prediction Error.**

I. Introduction

GAZE estimation technology plays a pivotal role in modern computer vision and human-computer interaction by tracking and interpreting the direction of a user's gaze[1]. This technology finds extensive applications in fields such as virtual reality (VR), augmented reality (AR)[5], driver assistance systems, medical diagnostics, and biometrics[2]. However, its accuracy is significantly compromised in low-light environments, where poor lighting conditions degrade image quality and impede the precise extraction of ocular features, thereby limiting the effectiveness and application scope of gaze estimation systems.

To overcome these challenges, we introduce GazeREC-Net, an advanced deep learning-based gaze restoration technique specifically designed to enhance gaze estimation accuracy in low-light conditions. We commenced by processing the publicly available MPIIFaceGaze[4] and ColumbiaGaze datasets to simulate low-light environments, thereby creating specialized degraded datasets that augment the representativeness and complexity of our experimental conditions. Subsequently, we developed GazeREC-Net, a novel gaze restoration model that synergizes Fourier transform techniques with convolutional neural networks to restore image quality while preserving critical features essential for accurate gaze estimation.

To assess the efficacy of GazeREC-Net, we conducted a comprehensive series of experiments comparing gaze estimation accuracy before and after image restoration using both datasets. The results unequivocally demonstrate that GazeREC-Net significantly enhances gaze estimation accuracy on restored images, evidenced by a substantial reduction in average prediction error[22]. These findings validate the robustness and practical potential of GazeREC-Net in ameliorating gaze information under low-light conditions, thus offering promising avenues for the application of gaze estimation technology in complex lighting environments.

The contributions of this paper are as follows:

1) Innovative Application of Fourier Transform and Gaze Modules: This study introduces a pioneering approach by integrating a Fourier transform module (FFTBlock) with a gaze module to develop a novel gaze restoration technique. This method leverages the Fourier transform to extract image features in the frequency domain, subsequently converting them back to the spatial domain via the inverse Fourier transform. This dual-domain processing not only restores critical image details and features but also incorporates essential gaze information. The innovative combination of these modules significantly enhances overall image quality, facilitating the extraction of key information pertinent to gaze estimation. This improved data input leads to more accurate and reliable gaze estimation results.

2) Development of the GazeREC-Net Gaze Restoration Model: This paper introduces GazeREC-Net, an advanced deep learning model meticulously designed to enhance gaze information under low-light conditions. The model integrates the Fourier transform, convolutional neural networks, and fully connected layers into a sophisticated multi-level framework for feature extraction and image restoration. This innovative architecture enables the precise reconstruction of degraded image quality in low-light environments, ensuring that the restored images preserve the essential features necessary for accurate gaze estimation.

3) Enhanced Generalization Capability in Low-Light Conditions: To bolster the model's generalization capability under low-light conditions, we employed an array of data augmentation techniques to transform normally lit images into simulated low-light environments. These techniques encompass adjustments to brightness and contrast, as well as the introduction of random noise. The resultant augmented dataset faithfully replicates real-world low-light scenarios, thereby providing diverse inputs for both training and testing phases. This comprehensive approach ensures the model's robust performance across a spectrum of complex low-light

Manuscript received June 7, 2024; revised October 16, 2024. This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

Jiayin Ku is a postgraduate student at School of Computer Science and software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 1340157320@qq.com).

Li Wang is a professor of the College of Computer Science and Tech nology Liaoning, Anshan 114051, China. (Corresponding author, e-mail: wangli9966@ustl.edu.cn).

conditions.

4) Comparative Experiments to Validate Model Effectiveness: To substantiate the effectiveness of GazeREC-Net, we conducted a series of systematic comparative experiments. Images were inputted into multiple gaze estimation models both before and after restoration. The experimental results consistently demonstrated a significant reduction in gaze estimation prediction error for the restored images. These findings underscore the efficacy of the gaze restoration model, indicating that our method markedly enhances gaze estimation accuracy in low-light conditions.

II. Related Work

In recent years, gaze estimation has emerged as a prominent research direction in computer vision, with broad applications in intelligent surveillance, virtual reality, and human-computer interaction. However, traditional gaze estimation methods often struggle to maintain sufficient accuracy despite inadequate lighting, noise interference, and blurred gaze features, presenting significant challenges for practical applications and underscoring the need for new methodologies.

Traditional gaze estimation methods primarily rely on geometric models, feature matching, and optical flow analysis. Some studies directly use facial images as input, employing convolutional neural networks (CNNs) to automatically extract deep features. These approaches have shown better performance than traditional methods using only eye images. However, facial images often contain redundant information, prompting researchers to filter out irrelevant features. For instance, Zhang et al. [4] proposed a spatial weighting mechanism that effectively encodes facial position into the CNN architecture by learning spatial weights through the activation maps of convolutional layers. This method predicts gaze direction in the camera coordinate system through an end-to-end learning strategy.

Recent research has also explored different network architectures, such as Transformers. Cheng et al. [6] were the first to investigate the performance of Transformers in gaze estimation, considering both pure and hybrid Transformer forms. The hybrid Transformer first generates feature maps through a CNN, which are then processed by the encoder in the Transformer.

In the context of single-stage gaze estimation, Zhang et al. proposed GazeOnce [7], the first single-stage method for estimating multiple human gaze directions through a single mapping. Other studies have opted to crop eye images from facial images and input them directly into the network for learning [8]. These methods typically employ a three-stream network to extract features from the face, left eye, and right eye images. Additionally, Cheng et al. [9] proposed a coarse-to-fine gaze estimation method, which first extracts features from facial images using a CNN and then refines the results using eye features.

Some research has focused on improving the design of convolutional layers. For example, Vieira [10] utilized attention-enhanced convolutional layers (AACoNv), producing more accurate results than ordinary convolutional networks. Biswas [11] introduced two novel techniques, I2D-Net and AGE-Net, which improve gaze estimation accuracy by eliminating standard features of participants' left and right eyes or assigning weights to features using an attention mechanism. Abdelrahman [12] proposed a robust CNN-based model, L2CS-Net, for predicting 3D gaze directions in unconstrained environments, achieving the lowest recorded angular error.

Despite the significant advances made by deep learning methods in gaze estimation, several technical challenges remain under low-light conditions. Issues such as uneven lighting, noise interference, model complexity, and training costs are key challenges that researchers need to address.

The model developed in this study enhances image visibility in low-light environments and accurately restores critical gaze-related information. By introducing specially designed Fourier transform-enhanced modules and deep feature fusion techniques, our method comprehensively extracts and optimizes gaze data during the image restoration process, significantly improving the accuracy and reliability of gaze estimation algorithms. This capability for precise gaze information restoration provides an innovative solution for low-light gaze estimation, demonstrating the potential of deep learning in complex visual tasks.

III. Methods

A. Model Introduction

We have proposed an enhanced method for gaze estimation under non-ideal conditions, aimed at solving the problem of inaccurate gaze estimation under low-light conditions. This technology covers the complete workflow from feature extraction to image reconstruction.

Feature extraction constitutes the initial stage of the image restoration process. This module processes gaze images under non-ideal conditions, capturing key visual features to improve the accuracy of gaze estimation. During the image restoration process, these key features are used to optimize gaze prediction results. The extracted features include the position, shape, and angle of the eye area, and involve data related to the eyes and pupils, ensuring that there is sufficient information for accurate processing during the image reconstruction stage.

Our method employs the HinBlock[13] module. HinBlock has evolved from the ResBlock, as shown in Figure 1(a). The ResBlock [14] module achieves feature extraction and nonlinear transformation through multiple layers of 3x3 convolution layers and LeakyReLU activation functions. HinBlock has been improved upon this basis by adding a 1×1 convolution layer to achieve residual connections for global feature extraction. Additionally, the HinBlock module adds instance normalization (IN) after the convolution layers to enhance processing stability. The HinBlock module first uses a 1×1 convolution layer to achieve residual connections, extracting global feature information. Subsequently, it processes through multiple 3x3 convolution layers and LeakyReLU activation functions, maintaining the size of the feature map while enhancing the network's nonlinear processing capability. The module continues through two layers of 3×3 convolution layers for local information extraction and nonlinear transformation processing. The HinBlock structure is shown in Figure 1(b).

After feature extraction, the image reconstruction module generates an enhanced gaze image using the extracted features. This process includes reconstructing the features into

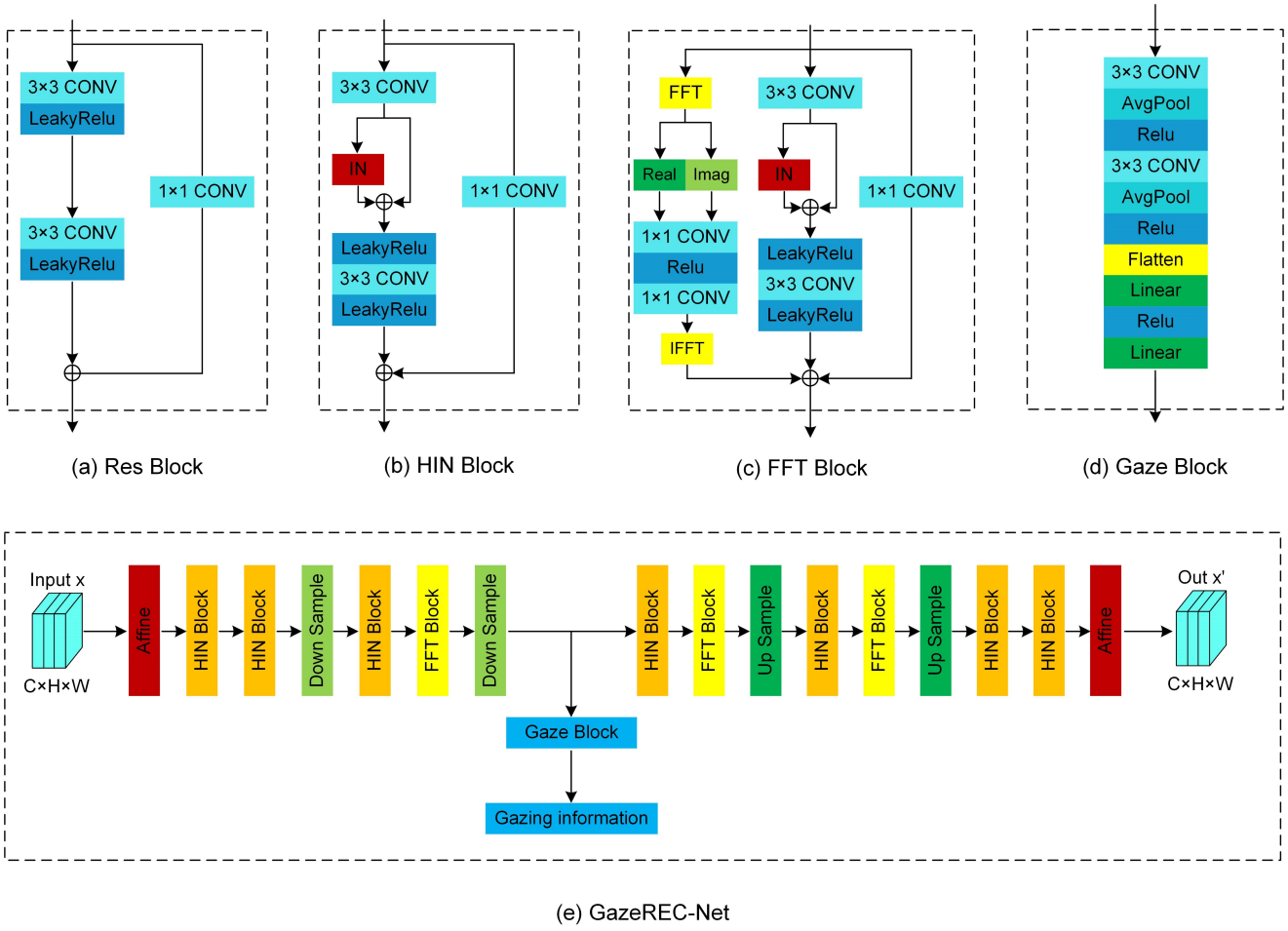

Fig. 1. The structure of different modules used in GazeREC-Net: (a) Res Block, (b) HIN Block, (c) FFT Block, (d) Gaze Block, (e) The overall architecture of GazeREC-Net.

an image, restoring obscured parts, and adjusting lighting conditions to make the image clearer, ensuring that the output image is visually close to the ideal state. We developed the FFT Block module, as shown in Figure 1(c). The innovation of this module lies in the use of Fourier transform technology, which significantly enhances the model's capability to process features in the frequency domain, further optimizing the accuracy of gaze estimation.

To further enhance the accuracy of gaze estimation, we designed a specialized Gaze Block module for processing information related to gaze estimation, as shown in Figure 1(d). This module can extract and integrate gaze information, making the final gaze estimation results more accurate.

The overall network architecture of GazeREC-Net is illustrated in the schematic shown as part (e) of Figure 1. It depicts the transformation process from input image to output image, featuring multiple key modules. Each module within the network is tasked with processing specific types of data essential for the system's operation.

Input and Output: The input image x undergoes a series of transformations to produce the modified output image x', both having the same channel count (C) and dimensions (HxW).

HinBlock Module: Responsible for advanced spatial feature extraction. Multiple HinBlock modules are distributed at different stages of the network to enhance the expressiveness of features and extract more complex spatial information.

FFTBlock Module: These modules use Fourier transforms to analyze and enhance the frequency domain characteristics of images. FFTBlock modules are called upon at different stages of image processing to fully utilize frequency domain features.

Upsampling and Downsampling: Steps following the FFTBlock that adjust the resolution of the feature maps to suit the different processing levels of the network.

Gaze Block: This module is specifically designed to handle information related to gaze estimation. It extracts crucial gaze information from the features extracted by the network and uses this information for final gaze estimation.

In the above text, we have preliminarily introduced the overall network architecture of the gaze estimation enhancement method, detailing the transformation process from input image to output image, and covering the roles and collaboration of multiple key modules. Next, we will delve into the specific implementations and technical details of the feature extraction module.

B. *Feature Extraction Module*

The core aim of the feature extraction module is to extract vital information from gaze images obtained under suboptimal conditions, which is crucial for subsequent tasks of image reconstruction and gaze prediction. This module uniquely combines traditional deep feature extraction techniques with

Fourier feature extraction, enhancing the system's capability to recognize and analyze the ocular region effectively.

The feature extraction module comprises two integral parts: deep feature extraction and Fourier feature extraction. These components work in tandem to enhance the representation of ocular features by integrating spatial and frequency domain information.

Deep Feature Extraction is primarily facilitated through the HinBlock (Hierarchical Interaction Block), designed to capture spatial information and local details effectively. The architecture of HinBlock incorporates several key technologies. Residual connections use 1×1 convolution layers to create pathways that enhance the network's training stability and feature transmission capability. This configuration aids in maintaining effective gradient flow in deep networks, thereby preventing gradient vanishing issues during training. A multi-layer convolution structure employs multiple 3×3 convolution layers interspersed with LeakyReLU activation functions, which not only preserves the dimensions of the feature maps but also enhances the network's capacity for nonlinear expression, allowing for detailed capture of local nuances and edge details. Instance normalization, applied after each convolution layer, standardizes feature maps for each instance, boosting the model's robustness and adaptability across varied imaging conditions.

Fourier Feature Extraction introduces an innovative aspect with the development of the FFTBlock, enabled by the application of the Fourier Transform. This allows the model to harness and utilize the frequency characteristics of images effectively. Under low-light conditions, images typically suffer from issues like noise and reduced contrast, which severely affect the accuracy of gaze estimation. By applying Fourier transform, GazeREC-Net converts images into the frequency domain where it independently analyzes the intensity and phase of each frequency component. This process helps distinguish noise-induced components from those essential for detailing. By enhancing significant high-frequency components and suppressing noise-related frequencies, the FFTBlock significantly improves image clarity and contrast. Furthermore, the accuracy of gaze estimation relies heavily on the quality of the ocular region's images. Through Fourier transform, GazeREC-Net optimizes features in the frequency domain, allowing not only for image quality enhancement but also for more precise localization and identification of crucial ocular features such as pupil positioning and eye contours. After processing the frequency domain features, the model converts these optimized features back to the spatial domain through the inverse Fourier transform, which restores the visual quality of the image and ensures accurate reconstruction of ocular features. Integrating these features with spatial domain deep features significantly boosts the overall accuracy of gaze predictions.

These enhancements confirm that the feature extraction module not only improves the capability to detect subtle changes in the image but also significantly elevates the accuracy of gaze estimation under challenging lighting and environmental conditions. By merging deep learning with signal processing techniques, our model demonstrates the potential of these combined technologies, forging new avenues for further research and application in gaze tracking technology.

C. Image Reconstruction Module

The main goal of the image reconstruction module is to use depth and Fourier features extracted from the feature extraction module to recreate high-quality gaze images. This involves enhancing visual details and adjusting lighting conditions to ensure the output images are both practical and visually appealing. Technical Details The image reconstruction process involves several key steps:

Feature Fusion: In GazeREC-Net's reconstruction module, feature fusion is critical. It combines depth features (spatial features extracted by HinBlock) and frequency features (extracted from the image's frequency components by FFTBlock) into a high-quality image. This is done by a weighted sum that not only simplifies feature dimensions but also enhances feature combination through optimized weights, improving the quality and details of the reconstructed image.

Channel Fusion: This is carried out through 1×1 convolution layers that help reduce computational complexity while preserving essential information. Residual connections are also used at each output stage to maintain deep feature transfer and learning efficiency. To restore the image's spatial resolution step-by-step, the module employs multiple levels of progressive upsampling, each followed by 3×3 convolution layers to refine features, focusing on enhancing details in critical areas like the eye region.

Lighting Adjustment: A dynamic adjustment module automatically analyzes and adjusts image brightness and contrast to suit complex lighting conditions and enhance visual appearance. Advanced image repair techniques in the final stage address any occlusions or damage, ensuring the integrity and usability of the output image. These carefully designed steps allow GazeREC-Net to recover high-quality, accurate gaze information from low-light images, significantly improving the precision and reliability of gaze estimation and supporting the development of low-light gaze tracking technology.

Upsampling and Feature Refinement: The reconstruction process uses progressive upsampling to gradually restore the image's spatial resolution. Each upsampling step is followed by convolution layers activated by ReLU to refine features and enhance image details. Upsampling increases spatial dimensions in the feature map to restore the image to its original size or a higher resolution. In GazeREC-Net, transposed convolution is used for upsampling because it effectively inserts spatial dimensions to restore image details. After each upsampling, feature refinement is necessary to enhance image details and clarity, involving multiple layers of 3×3 convolution that introduce non-linearity with LeakyReLU activation functions, enhancing the model's ability to capture complex image features.

Lighting Adjustment and Optimization: The lighting adjustment and optimization module in GazeREC-Net improves brightness and contrast of images under low lighting, enhancing gaze estimation accuracy and visual comfort. It uses Fourier transformation to analyze image frequencies, allowing automatic adjustments based on lighting conditions.

Loss Function Design: To optimize image reconstruction quality and gaze prediction accuracy, we use a composite loss function that includes L1 loss and cross-entropy loss. This allows simultaneous image reconstruction and gaze direction classification.

L1 loss, or least absolute deviation, is used primarily for image reconstruction, calculated as:

$$L_{L1} = \frac{1}{N}\Sigma_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

Here, y_i represents the pixel values in the target image, $\hat{y}_i$ are the corresponding pixel values in the predicted image, and N is the total number of pixels in the image.

Cross-Entropy Loss Used for evaluating gaze classification effectiveness, it is expressed as:

$$L_{\mathrm{CE}} = -\sum_{c=1}^{M} y_{o,c}\log\left(p_{o,c}\right) \tag{2}$$

Here, $y_{o,c}$ is an indicator variable that denotes whether class label c is the correct classification for sample o while $p_{o,c}$ is the probability predicted by the model that sample o belongs to class c.

The total loss of the model is a weighted sum of the two aforementioned losses, represented as:

$$L_{\mathrm{total}} = \alpha L_{\mathrm{L1}} + \beta L_{\mathrm{CE}} \tag{3}$$

where α and β are the weight coefficients that adjust the contribution of the two parts of the loss. Through this process, the image reconstruction module not only restores the physical attributes of the image, such as texture and edges, but also enhances the visual quality of the image, improving its performance under complex lighting conditions. Furthermore, the reconstructed image displays higher precision and reliability in gaze estimation tasks, demonstrating the effectiveness of the proposed method.

D. FFTBlock

Fourier transform[15] is a crucial signal processing technique that converts a signal from the time domain to the frequency domain, allowing for the analysis of the presence and contribution of different frequency components within the signal. In this study, the pre-processed eye movement image data undergoes a Fourier transform. This transform converts images from the spatial domain to the frequency domain, transforming them into spectral images. Through Fourier transformation, spatial frequency information, including high and low frequency components, is extracted from the images. We utilize the Fourier transform to convert eye movement images from the time domain to the frequency domain, to capture spatial frequency information within the eye movement images, and combine this with convolutional neural networks for gaze estimation.

For a one-dimensional continuous signal $f(x)$, its Fourier transform $F(u)$ is defined as follows:

$$F(u) = \int_{-\infty}^{\infty} f(x)\cdot \mathrm{e}^{-i2\pi x}\mathrm{d}x \tag{4}$$

where u represents the frequency variable, and i denotes the imaginary unit. This transformation decomposes the time-domain signal $f(x)$ into a linear combination of complex exponential functions of infinite frequency components, facilitating the signal's conversion from the time domain to the frequency domain. This process not only reveals the presence and amplitude of frequency components within the signal but also decodes their phase information, making it an indispensable tool in modern signal processing.

For a two-dimensional discrete signal $f(x, y)$, its discrete Fourier transform $F(u, v)$ is defined as follows:

$$F(u,v) = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cdot e^{-i2\pi\left(\frac{ux}{N}+\frac{vy}{N}\right)} \tag{5}$$

Here, N represents the number of pixels in the horizontal and vertical directions of the image, x,y are the spatial coordinates of the image, and u,v are the coordinates in the frequency domain, corresponding to the frequencies in the horizontal and vertical directions, respectively. Through this transformation, the two-dimensional image is converted to the frequency domain, where the complex value at each frequency point (u, v) reflects the amplitude and phase information of that frequency component in the image, allowing us to analyze the frequency distribution and its spatial characteristics within the image.

The discrete Fourier transform converts a two-dimensional image into a complex representation in the frequency domain, where the amplitude and phase information of each frequency reflects the distribution of different spatial frequencies within the image. The two-dimensional discrete Fourier transform (DFT) allows us to identify which areas of the image contain high or low-frequency components and their positions across the entire image. In the frequency domain, we extract frequency domain features from the spectrum $F(u, v)$ obtained by the two-dimensional discrete Fourier transform. This step is aimed at extracting meaningful features from the spectrum that indicate the importance and distribution of different spatial frequency components in eye movement images for subsequent gaze position estimation.

After performing the two-dimensional discrete Fourier transform, we calculate the spectral energy $E(u, v)$ to assess the energy intensity of each frequency component in the image, using the formula:

$$E(u,v) = \mid F(u,v) \mid^2 \tag{6}$$

Here, F(u,v) is the result of the two-dimensional Fourier transform of the eye movement image, and $\mid F(u,v) \mid^2$ represents the magnitude at the frequency point (u, v), which is the intensity of that frequency component in the spectrum. By calculating spectral energy, we can reveal the distribution and significance of different spatial frequency components in the eye movement image.

To gain a more comprehensive understanding of the image's frequency characteristics, we also calculated the average spectral energy, which represents the energy distribution across the entire spectrum. The average spectral energy is obtained by averaging the spectral energy $E(u, v)$ within the frequency domain as follows:

$$E = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} E(u,v) \tag{7}$$

Here, M and N represent the width and height of the spectrum, indicating the range of frequencies within the frequency domain. $E(u, v)$ is the energy at the coordinate (u, v) on the spectrum. The average spectral energy reflects the combined intensity of frequency components throughout

the frequency domain, providing a valuable perspective for understanding and analyzing eye movement images.

To eliminate scale differences between different frequency components, we normalized the spectral energy. Specifically, the spectral energy $E(u, v)$ was normalized using the following formula:

$$\tilde{E}(u, v) = \frac{E(u, v)}{\max\big(E(u, v)\big)} \tag{8}$$

where $E(u, v)$ is the original spectral energy and $\max\big(E(u, v)\big)$ is the maximum energy value in the spectrum.

This normalization process ensures that all spectral energy values are within the range of 0 to 1, facilitating subsequent comparison, visualization, and analysis. This enhancement not only deepened our understanding of the image's frequency domain features but also provided data support for subsequent gaze position prediction.

E. GazeBlock

To ensure the model can accurately extract and predict gaze-related information during training, this study designed a specialized network block—GazeBlock. GazeBlock optimizes the precise prediction of gaze direction by processing fine spatial and frequency features. It utilizes a combination of convolutional layers and pooling layers, not only compressing feature representations to meet the demands of gaze prediction but also enhancing the model's sensitivity to gaze information through a structured feature extraction process. Additionally, the design of GazeBlock focuses attention on key features of gaze prediction during training, ensuring the learning of effective features directly related to gaze estimation tasks.

GazeBlock is specifically designed for gaze estimation tasks, comprising deep convolutional network layers and sequential pooling layers. This configuration helps reduce the spatial dimensions of features while retaining crucial information essential for predicting gaze direction. In the feature extraction phase, features are extracted using 3x3 convolution kernels and ReLU activation functions, ensuring the effectiveness of nonlinear processing. Subsequent average pooling and adaptive pooling steps further reduce the spatial dimensions of the feature maps, which not only helps to alleviate computational burdens but also maintains the stability of the model's output.

After obtaining compressed and information-rich featuresMA, GazeBlock processes these features through a sequence of fully connected layers. Initially, a Flatten operation transforms the multidimensional feature maps into a one-dimensional vector, which is then finely captured by two linear transformation layers that are critical for predicting gaze direction. These layers not only map[18] high-dimensional features to predictive outputs through linear transformations but also enhance the model's non-linear capability to capture complex feature relationships through activation functions.

By integrating GazeBlock, this study ensures the model's efficiency and focus in the feature extraction and transformation process, thereby significantly enhancing the accuracy of gaze estimation. Further explanation is provided on the application of GazeBlock in gaze estimation tasks, showcasing its specific configuration and process within the network.

IV. Experiment

A. Dataset Preparation and Preprocessing

In this study, we utilized both the ColumbiaGaze and MPIIFaceGaze datasets for our gaze estimation experiments, as illustrated in Figure 2. The ColumbiaGaze dataset comprises 5880 full-face images from 56 subjects, encompassing a variety of lighting conditions and head poses. The MPIIFaceGaze dataset consists of 45000 facial images with gaze annotations from 15 subjects, captured over several months under diverse lighting conditions. Together, these datasets provide a comprehensive collection of images, ensuring diversity and broad environmental coverage.

To simulate low-light environments, we generated a series of low-light images by blending standard lighting images with completely black background images at random ratios and adjusting the brightness and contrast randomly, as shown in Figure 3. Furthermore, we introduced varying levels of random noise to enhance the robustness of our model.

These preprocessing steps not only improved the model's performance under challenging lighting conditions but also have practical implications for applications such as intelligent surveillance and virtual reality[17].

Fig. 2. Sample facial images from the MPIIFaceGaze dataset under standard lighting conditions.

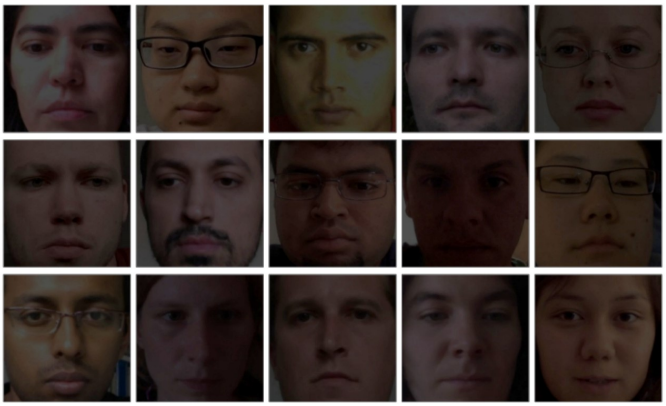

Fig. 3. Facial images generated to simulate low-light conditions, showing varying levels of brightness adjustments and random noise.

B. Model Training

The GazeREC-Net model was developed using the PyTorch framework and was trained and tested on an NVIDIA

GeForce RTX 4090 GPU. The experimental setup included a 12-core CPU and 29GB of RAM for efficient data processing. The training process employed a batch size of 32, spanned 100 epochs, and utilized the Adam[21] optimizer with a learning rate set at 0.0001. The input images were resized to 224×224 pixels and processed in the RGB three-channel format.

To evaluate the performance of GazeREC-Net under low-light conditions, we initially selected subsets of images from the ColumbiaGaze and MPIIFaceGaze datasets for preliminary validation, ensuring that the sample encompassed various lighting and environmental conditions. These images were algorithmically processed to simulate low-light environments by blending standard lighting images with completely black background images at random ratios and adjusting their brightness and contrast. Additionally, varying levels of random noise were introduced to enhance the model's robustness. The original images were retained for comparative analysis.

The experiment was expanded to include the entire dataset following the preliminary validation. The leave-one-out cross-validation (LOOCV) method was employed to rigorously train and evaluate the GazeREC-Net model. In each iteration of LOOCV, one subject was excluded from the training set and used as the test set, with the remaining subjects comprising the training set[20].

Experimental results demonstrated that GazeREC-Net significantly outperformed comparison models regarding Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), highlighting its superior image restoration capabilities. Additionally, the gaze estimation accuracy was markedly improved, with errors in the restored images substantially lower than those in the degraded images. These findings confirm the model's effectiveness and potential for practical applications in low-light conditions.

C. Experimental Results and Analysis

1) Evaluation of Image Restoration: This study assessed image restoration quality using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR measures image restoration quality by comparing the error between the maximum possible and actual pixel values in the original and restored images. A higher PSNR indicates less distortion and higher restoration quality. SSIM quantifies the visual effects of images by simulating the human eye's ability to perceive structural information, assessing the similarity between two images. An SSIM value close to 1 typically indicates that the restored image is visually similar to the original image. These metrics provide an objective assessment of the quality transition from a degraded to a restored state, enabling us to quantitatively analyze the performance of GazeREC-Net in image restoration.

CFTNet primarily improves image quality by manipulating image features within the frequency domain. This model transforms images from the spatial to the frequency domain via the Fourier transform and independently processes the real and imaginary parts to enhance specific frequency components. However, CFTNet fails to effectively preserve critical visual information during the image reconstruction process, particularly under low-light conditions, where its performance is limited by the constraints of frequency domain analysis.

HINet employs a hierarchical network structure that enhances image detail through deep spatial feature interactions. Each HinBlock in HINet includes residual connections and instance normalization, which help to improve feature transmission and network training stability. Although HINet performs excellently in extracting spatial domain features, its ability to restore subtle features in low-light images remains limited.

In contrast, GazeREC-Net combines advanced techniques from Fourier transform and deep learning, optimizing feature extraction in frequency and spatial domains. It explicitly addresses gaze estimation issues under low-light conditions through an innovative network architecture. By finely tuning image features in both frequency and spatial domains, GazeREC-Net significantly improves gaze estimation accuracy and notably enhances the visual quality of images.

TABLE I
IMAGE RECOVERY METRICS COMPARISON

Method	MPIIFaceGaze PSNR	MPIIFaceGaze SSIM	ColumbiaGaze PSNR	ColumbiaGaze SSIM
CFTNet	20.5	0.65	21.0	0.67
FFTNet	29.3	0.84	29.8	0.85
HINet	30.1	0.87	30.5	0.88
GazeREC-Net	32.4	0.92	32.8	0.93

This table compares the performance of CFTNet, FFTNet, HINet, and GazeREC-Net using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). GazeREC-Net shows the highest values, indicating its superior image recovery quality.

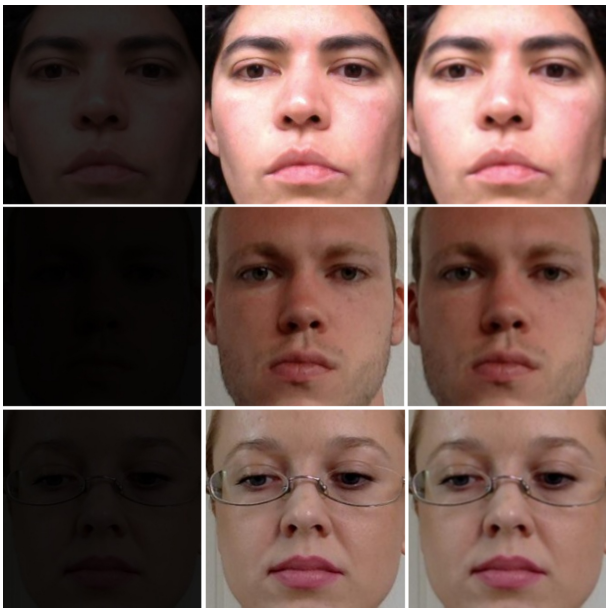

Fig. 4. Comparison of Gaze Recovery Results.

The Figure 4 showcases the images of three selected individuals, demonstrating GazeREC-Net's significant enhancement in image quality through a clear visual comparison across the rows. Each row corresponds to one individual, with the first column displaying the low-light image, the second column the original, and the third column the recovered image.

TABLE II
GAZE ESTIMATION ERROR COMPARISON ON MPIIFACEGAZE AND COLUMBIAGAZE

	MPIIFaceGaze					ColumbiaGaze				
Method	Original Image	Unrecovered Image	CFT	HINet	Ours	Original Image	Unrecovered Image	CFT	HINet	Ours
Gaze-TR	4.17°	4.92°	4.53°	4.27°	4.06°	5.01°	5.62°	5.43°	5.10°	5.05°
Dilated-Net	5.21°	5.32°	5.13°	4.96°	4.83°	5.43°	6.02°	5.67°	5.40°	5.35°
CA-Net	4.12°	4.96°	4.68°	4.38°	4.12°	5.05°	5.65°	5.30°	5.05°	5.00°
L2CS-Net	3.92°	4.85°	4.34°	4.19°	3.92°	4.90°	5.93°	5.15°	4.90°	4.35°
MTGLS	4.23°	4.43°	4.38°	4.35°	4.23°	4.50°	5.10°	4.75°	4.50°	4.47°

2) Gaze Estimation Accuracy Analysis: To comprehensively assess the effectiveness of GazeREC-Net in enhancing gaze estimation accuracy, we extended our experimental

In this section, we focus on demonstrating the image recovery capabilities of GazeREC-Net under low-light conditions. Three representative individuals from the MPIIFaceGaze dataset were selected for an in-depth analysis to illustrate the comparative effects before and after model recovery.

Scope to include all subjects from both the ColumbiaGaze and MPIIFaceGaze datasets. The selection of these participants considered diversity, including different genders, whether glasses were worn, and variations in head posture[23]. This diverse setup helps to test and verify the performance of GazeREC-Net under various real-world conditions with insufficient lighting.

Our experiments encompassed several existing gaze estimation models, including Gaze-TR[6], Dilated-Net[16], CA-Net[9], L2CS-Net[17], and MTGLS[19]. We compared gaze estimation errors before and after applying our GazeREC-Net technology. Table II compares gaze estimation errors across different recovery technologies, including the original data without processing, data recovered using the CFTNet and HINet models, and data recovered using GazeREC-Net.

In this study, we comprehensively assessed the impact of GazeREC-Net technology on gaze estimation accuracy. Specifically, the GazeTR model was evaluated using the MPIIFaceGaze dataset, which included 15 participants. The selection of these participants ensured extensive coverage of various lighting conditions, whether participants wore glasses, and variations in head posture, thus providing general applicability and reliability of the test results.

Our analysis focused on comparing the gaze estimation errors of each participant under three different conditions: baseline (unprocessed original data), deteriorated (under low-light conditions where the quality of gaze data deteriorated), and post-recovery using GazeREC-Net.In Table III, the first column lists 15 different subjects, the second column represents the error in the baseline state, the third column indicates the error after deterioration, and the fourth column shows the error after recovery.

Results indicated that for most participants, the gaze estimation error after recovery with GazeREC-Net was significantly lower than in the deteriorated state and, in most cases, was close to or better than the baseline state. This finding demonstrates the effectiveness of GazeREC-Net in restoring gaze data under low-light conditions and highlights its potential to enhance gaze estimation accuracy.

In our study, we validated the universality and effectiveness of GazeREC-Net technology by comparing the average

TABLE III
GAZE ERROR ACROSS DIFFERENT SUBJECTS

Subject	Image Before Recovery	Image After Recovery	Original Image
P00	2.93	2.24	2.19
P01	4.31	4.03	4.19
P02	6.89	6.02	5.99
P03	3.09	2.60	2.51
P04	3.57	2.93	2.55
P05	4.40	3.78	3.72
P06	5.05	3.48	3.40
P07	4.93	4.09	3.91
P08	5.18	4.58	4.24
P09	5.45	3.35	3.93
P10	5.24	4.65	4.90
P11	5.46	4.43	4.30
P12	4.73	4.06	4.14
P13	5.67	4.75	4.71
P14	6.86	5.98	6.08
Avg	4.92	4.06	4.05

gaze estimation errors across all participants from both the MPIIFaceGaze and ColumbiaGaze datasets. The selection of participants considered diverse conditions, including different genders, whether participants wore glasses, and variations in head posture, ensuring general applicability and reliability of the results. This comprehensive analysis highlights the practical application value of our gaze restoration technology and its potential for enhancing gaze estimation in low-light and other visually challenging environments[22].

Our approach involved using GazeREC-Net for image restoration and applying the Gaze-TR model to predict gaze angles. We compared the gaze estimation errors of each participant under three different conditions: baseline (original unprocessed data), deteriorated (low-light conditions without recovery), and post-recovery (using GazeREC-Net for image restoration). In the bar chart, gray bars represent the error in the baseline state, orange bars show the error after deterioration and blue bars indicate the error after recovery.

Furthermore, detailed ablation studies were conducted to highlight the synergistic effects of the HIN, Gaze, and FFT blocks within the GazeREC-Net architecture. By isolating and combining these components, we identified their individual and combined contributions to gaze estimation accuracy. When all three elements were activated, the gaze estimation errors reached their minimum at 4.06° for the MPIIFaceGaze dataset and 4.10° for the ColumbiaGaze dataset, emphasizing the critical role of multi-component synergy in achieving optimal performance. Table IV displays the results of the ablation experiments, showcasing each component's contribution.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS

Hinet	gazeblock	FFTblock	MPIIFaceGaze Gaze Estimation Error	ColumbiaGaze Gaze Estimation Error
✓			4.21°	4.23°
✓	✓		4.17°	4.20°
✓		✓	4.13°	4.15°
✓	✓	✓	4.06°	4.10°

V. CONCLUSION

This research introduced the GazeREC-Net model, which effectively enhances gaze estimation accuracy under low-light conditions by integrating advanced Fourier transform and deep learning techniques. Our extensive evaluations, conducted on both the MPIIFaceGaze and ColumbiaGaze datasets, have demonstrated significant improvements in the quality of low-light images and the precision of gaze estimation.

The results conclusively demonstrate GazeREC-Net's capability to accurately restore gaze information in challenging lighting environments, substantially reducing average gaze prediction errors. This restoration proves the model's adaptability to real-world scenarios, often characterized by variable lighting conditions, and underscores its robustness across different datasets. Furthermore, the innovative incorporation of the Fourier Transform module (FFTBlock) and the gaze-specific processing module (GazeBlock) introduces new methodologies for advancing gaze estimation technology under complex lighting conditions.

Ultimately, GazeREC-Net not only offers a novel method for gaze estimation within the field of computer vision but also opens new possibilities for image processing and feature recovery in low-light conditions. We anticipate that the insights gained from this study will significantly benefit the practical application of gaze-tracking technologies across diverse sectors, including autonomous driving, intelligent monitoring, and virtual reality. This study paves the way for broader adoption and technological advancements.

REFERENCES

[1] R. M. Rahal and S. Fiedler, "Understanding cognitive and affective mechanisms in social psychology through eye-tracking," *Journal of Experimental Social Psychology*, vol. 85, pp. 103842, 2019.

[2] Y. Cheng, X. Zhang, F. Lu and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Transactions on Image Processing*, vol. 29, pp. 5259–5272, 2020.

[3] S. Park, A. Spurr, O. Hilliges, "Deep pictorial gaze estimation," in *Proceedings of the European conference on computer vision (ECCV) 2020*, vol. 29, pp. 5259–5272.

[4] X. Zhang, Y. Sugano, M. Fritz and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops 2017*, pp. 51-60.

[5] J. Ratcliffe, F. Soave, N. Bryan-Kinns, L. Tokarchuk and I. Farkhatdinov, "Extended reality (XR) remote research: A survey of drawbacks and opportunities," in *Proceedings of the 2021 CHI conference on human factors in computing systems 2017*, pp. 51-60.

[6] Y. Cheng and F. Lu, "Gaze estimation using transformer," in *2022 26th International Conference on Pattern Recognition (ICPR). IEEE 2022*, pp. 3341-3347.

[7] M. Zhang, Y. Liu and F. Lu, "Gazeonce: Real-time multi-person gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022*, pp. 4197-4206.

[8] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2016*, pp. 2176-2184.

[9] Y. Cheng, S. Huang, F. Wang, C. Qian and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence 2020*, vol. 34, no. 7, pp. 10623-10630.

[10] G. L. Vieira and L. Oliveira, "Gaze estimation via self-attention augmented convolutions," in *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE 2021*, pp. 49-56.

[11] P. Biswas, "Appearance-based gaze estimation using attention and difference mechanism," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021*, pp. 3143-3152.

[12] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi and L. Dinges, "L2cs-net: Fine-grained gaze estimation in unconstrained environments," in *2023 8th International Conference on Frontiers of Signal Processing (ICFSP). IEEE 2023*, pp. 98-102.

[13] L. Chen, X. Lu, J. Zhang, X. Chu and C. Chen, "Hinet: Half instance normalization network for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021*, pp. 182-192.

[14] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2016*, pp. 770-778.

[15] M. Iqbal, M. M. Riaz, A. Ghafoor, A. Ahmad and S. S. Ali, "Out of focus multi-spectral image de-blurring using texture extraction and modified fourier transform," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12671-12684.

[16] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Asian Conference on Computer Vision. Cham: Springer International Publishing 2018*, pp. 309-324.

[17] Y. Sugano, Y. Matsushita and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2014*, pp. 1821–1828.

[18] P. Kettunen and J. Oksanen, "Efects of Unsupervised Participation overthe Internet on a Usability Study about Map Animation," in *New Directions inGeovisual Analytics: Visualization, Computation, and Evaluation 2018*, pp. 1–7.

[19] S. Ghosh, M. Hayat, A. Dhall and J. Knibbe, "Mtgls: Multi-task gaze estimation with limited supervision," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision 2022*, pp. 3223-3234.

[20] X. Zhang, Y. Sugano and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. International Symposium on Eye Tracking Research and Applications (ETRA) 2018*, pp. 1–9.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

[22] Y. Zhang, M. K. Chong, J. Müller, A. Bulling and H. Gellersen, "Eye tracking for public displays in the wild," *Personal and Ubiquitous Computing*, vol. 19, no. 5-6, pp. 967–981, 2015.

[23] S. Duffner and C. Garcia, "Visual focus of attention estimation with unsupervised incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2264–2272, 2016.