IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 22 October 2023, accepted 16 November 2023, date of publication 20 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335249

RESEARCH ARTICLE

Hybrid Swin Transformer-Based Classification of Gaze Target Regions

GONGPU WU[1], **CHANGYUAN WANG**[2], **LINA GAO**[1], **AND JINNA XUE**[2]
[1]School of Optoelectronic Engineering, Xi'an Technological University, Xi'an 710021, China
[2]School of Computer Science, Xi'an Technological University, Xi'an 710021, China

Corresponding author: Changyuan Wang (cyw901@163.com)

The work of Changyuan Wang was supported by the National Natural Science Foundation of China funded by the Chinese Government under Project 52072293.

ABSTRACT Inferring gaze targeting or gaze following is an effective approach for comprehending human behavior and intentions. This paper employs a non-intrusive appearance-based tracking technique, utilizing a binocular stereo vision camera to capture the face image and head pose to address errors caused by problems such as the disappearance of the eye image and head deflection occlusion in image capture. Each gaze direction is determined based on a single image frame. To improve the classification and detection of the gaze target region by effectively handling head motion and view direction, this paper proposes a hybrid structure for the Swin Transformer gaze target region classification method. The facial image features are extracted using both the ResNet50 model and the Swin Transformer model, followed by fusing head pose features to categorise the gaze target area. The study also compares the classification effects of various structural models. The analysis of the results demonstrates that the hybrid Swin Transformer model outperforms in classifying and detecting the gaze target region, achieving an accuracy rate of 90%. Finally, the research examines the gaze of flight trainees during flight missions by using a heatmap, which lays the groundwork for future analyses of pilot attention and operational intentions during flights.

INDEX TERMS Gaze estimation, swin transformer, computer vision, region classification.

I. INTRODUCTION

Due to the rapid development in the fields of computer vision and artificial intelligence, the success and widespread adoption of deep learning have greatly enhanced the performance of eye-tracking [1], [2]. Researchers' interest in understanding and simulating the human visual system is growing. Accurate classification and detection of gaze target regions [3] are vital to achieve highly intelligent automation systems in numerous application domains. This challenge involves identifying specific regions of interest within images and videos where an observer's focuses their attention. These regions usually contain critical information about the scene or objects captured in them. However, head movements are complex, environmental factors are unstable, and minor changes in viewing angles or head positions can cause significant alterations in the appearance or gaze direction of the eyes. In situations where severe head and eye rotations occur during image capture, techniques based on Convolutional Neural Networks (CNN) may be adversely affected [4], resulting in the loss of eye images during image capture. Existing methods for gaze target classification still have certain limitations when addressing these challenges.

Eye-tracking technology and gaze target area estimation are closely related, and commonly utilised in eye-tracking research and human-computer interaction. Eye-tracking technology [5] provides a means of collecting eye movement data, while gaze target area estimation [6] is a method for analyzing this data to obtain information about user visual attention. Combining gaze tracking technology with gaze target area

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

estimation technology can help us gain insights into a pilot's visual behavior in specific tasks or situations.

This study aims to introduce a non-invasive appearance-based tracking technique [7] for acquiring eye-tracking data and a novel gaze target area classification method based on a hybrid Swin Transformer to improve the accuracy of gaze target area classification and detection. The hybrid Swin Transformer is a deep learning model that inherits the residual structure of ResNet50 [8] and the local perception capability of Swin Transformer [9], organically combining the two. This model exhibits exceptional performance when processing image information, particularly in addressing changes in head posture and gaze direction, making it uniquely advantageous. In this paper, we conduct research using an appearance-based gaze tracking approach, utilizing a binocular stereoscopic vision system to acquire facial images and head postures. Feature extraction and fusion are performed on facial images and head postures obtained from different cameras, and the hybrid Swin Transformer model is employed for the classification of gaze target areas on the cockpit instruments of pilots. This research aims to improve pilot-human interaction efficiency and the accuracy of pilot gaze tracking while laying the foundation for further exploration of pilot attention mechanisms.

In this paper, we will delve into a discussion of the principles of non-intrusive appearance-based tracking technology and the hybrid Swin Transformer. We will also explore how to apply these principles to gaze target area classification, including the methods for integrating gaze target area estimation to determine the user's points of visual interest. This will aid in gaining a better understanding of user attention allocation in various tasks and scenarios. Furthermore, our research will focus on an appearance-based gaze-tracking approach, utilizing a binocular stereoscopic vision system to capture facial images and head postures. Additionally, we will introduce how to fuse facial image features and head posture information to enhance the performance of classification and detection. Lastly, we will present experimental results and performance analysis, demonstrating the exceptional performance of the hybrid Swin Transformer model in gaze target area classification tasks.

Our main contributions are as follows:

- We propose a hybrid model algorithm of ResNet50 and Swin Transformer, which is used to classify the gaze target area and construct a data set of the flight cadets ' gaze target area during the simulated flight.
- We compared and analyzed the classification results of the Swin Transformer, Vision Transformer, ResNet50 + Swin Transformer, and ResNet50 + Vision Transformer and the classification results with or without head posture.
- We propose to use the heat map to analyze the fixation of the flight cadets in the current flight scene, which paves the way for the subsequent pilot's attention evaluation.

The structure of this article is as follows:

- In the first section, we discuss the background and related work on non-intrusive appearance-based gaze tracking technology and gaze target area estimation;
- In the third section, we introduce the composition of the data set, the principles of the VIT model, ResNet model, and Swin Transformer model, as well as the gaze target area classification method based on the ResNet50 and Swin Transformer hybrid model;
- In the fourth section, we compare and analyze the classification effect of the model and the visual display of the experimental results;
- In the fifth section, we summarize the research conclusions and present future research ideas.

II. GUIDELINES FOR MANUSCRIPT PREPARATION

A. BASED ON NON-INVASIVE APPEARANCE GAZE TRACKING TECHNOLOGY

Gaze tracking technology is a technology aimed at understanding user intent and interests [10], focusing on the relationship between image data and gaze direction. Gaze tracking technology has a wide range of applications in various fields, including human-computer interaction [11], [12], medical diagnosis [13], defense and military [14], aviation safety [15], traffic safety [16], [17], [18], virtual reality [19], [20], biosecurity [21], and more. Gaze can be considered one of the behaviors that reflect human attention. Especially in the military domain, gaze tracking technology, as a crucial indicator for assessing pilot attention, holds significant importance. The study of pilots' gaze behavior is of great significance in this context.

Gaze-tracking technology has matured over several decades of research [22] and is broadly categorized into invasive and non-invasive systems. Intrusive systems are highly accurate but inconvenient. Non-invasive systems capture facial images through cameras and use computer vision and machine learning algorithms to estimate the direction of gaze, which is more comfortable and has broad prospects. Method research on gaze tracking technology is mainly divided into feature-based methods and appearance-based methods. The feature-based method [23] generally uses Purkin spots obtained by corneal reflection to detect pupils. This method establishes a line-of-sight point model. Although the error is small, it is too dependent on the light source and has a cumbersome calibration process. The appearance-based method [7] mainly uses human eyes or facial images as input, establishes a mapping model between features and gaze direction, and outputs the gaze direction. This method is simple in design, low in cost, and highly robust. We will consider using appearance-based methods for research in this article.

We will conduct a study from the perspective of non-invasive gaze tracking technology and review recent advances in non-invasive gaze tracking technology. Naqvi et al. [24] proposed a deep learning-based gaze detection method using near-infrared camera sensors that do not require initial user calibration. It extracts facial and dual-eye images to obtain gaze features. This method is used for driver gaze classification and exhibits good accuracy. However, errors may arise when head and eye rotation cause one eye to disappear. Yiu et al. [25] employed a Fully Convolutional Neural Network (FCNN) to segment the pupil region, improving gaze estimation accuracy, though lacking in head pose research. Cheng et al. [26] introduced the FARE-Net model, which uses facial and eye images to predict 3D gaze direction for both eyes, adapting strategies adaptively. However, its effectiveness is limited under low-light conditions and head rotation. Wang et al. [23] employed random forest regression to learn the mapping between deep features and gaze coordinates, achieving good performance but lacking real-time driver tracking capability. Sayeed et al. [27] proposed a system that can detect eye and head movement during reading, using pupil-iris detection for eye position and nose tip detection for head orientation. However, this system has issues with the precision of head and eye position. To address issues related to head pose differences and occlusion, Dai et al. [28] introduced a gaze-tracking method based on binocular feature fusion and convolutional neural networks. It uses left and right eye images and facial information as input, with experimental results showing superior performance with binocular feature fusion.

Cazzato et al. [29] discussed the significant advancements in the field of gaze tracking brought about by computer vision and machine learning. They proposed a new classification system and highlighted that computer vision has made gaze-tracking systems increasingly precise, but evaluating them using a single metric is quite challenging. In this paper, to mitigate errors caused by eye image disappearance and head deviation leading to occlusion, we first employed the approach presented by George et al. [30]. However, we did not extract left and right eye images separately; instead, we captured facial images using left and right cameras. Subsequently, we incorporated the binocular feature fusion method proposed by Dai et al. [28]. We further refined this approach and assessed gaze tracking using multiple metrics. Specifically, we introduced the fusion of left and right facial features with head pose features as inputs to our model. This model tracks the pilot's gaze and outputs the corresponding gaze target region number on the cockpit instruments.

B. ESTIMATION OF THE GAZE TARGET

Gaze target area estimation analyzes the gaze pattern of human eyes to determine the target area they focus on. This technology is mainly based on the working principle of the human visual system. By tracking eye movements and gaze points, it can infer the distribution of attention of the human eye when observing a scene. This method uses cameras or other sensors to track the position and movement of the eyes. It determines the gaze target area by analyzing features such as faces, pupils, or eyeballs.

In the context of estimating gaze targets, Chong et al. [31] introduced a multi-task learning approach and neural architecture. This approach explicitly represents gaze direction and processes gaze targets to estimate the general areas of visual attention in images by the general population. Recasens et al. [32] proposed a deep neural network-based gaze tracking method to predict objects that might be observed in a scene, achieving high estimation accuracy. While this method is also applicable to detecting gaze targets for multiple individuals, it is limited to cases where both the person and their gaze target appear in the same image. To address this limitation, Recasens et al. [33] introduced cross-frame gaze target detection for videos, allowing individuals and their gaze targets to appear in different video frames. We conduct research by collecting real-time face images of student pilots when they gaze at the target area during simulated flight and using a hybrid Swin Transformer model to estimate the gaze target area. Regarding driver attention, Hu et al. [34] used low-level features, static visual saliency maps, and dynamic optical flow information as input feature maps. They combined high-level semantic descriptions with gaze probability maps transformed from gaze directions to propose a data-driven, multi-resolution neural network suitable for estimating driver attention. Hu et al. [35] extended gaze target estimation from 2D images to 3D space to infer the driver's 3D gaze targets. They use head pose encoding, scene images, and facial images as inputs to generate predicted 3D gaze vectors and predicted gaze heatmaps as outputs. Personalized driver gaze area estimation systems have seen significant improvements, but a general gaze area estimation framework that is invariant to different subjects, viewpoints, and scales is still lacking. In the context of human-target interaction, Hu et al. [36] introduced interactive attention to investigate gaze target estimation. They utilized a visual-spatial map to analyze the interaction probabilities between individuals and targets within a scene. Chakraborty et al. [6], following their introduction of image segmentation and facial feature detection techniques, employed deep learning architectures to analyze eye movements and gaze estimation. They then applied these models to robotic systems for the calculation of visual attention for tasks such as reading, browsing, and writing.

In their study, Vora et al. [16] employed Convolutional Neural Networks to classify gaze areas from different subjects and various viewpoints, comparing the effectiveness of using different facial image components as input strategies. In contrast, in our research, we didn't rely on scene images to obtain head pose information but rather computed head pose through mathematical calculations. We divide the cockpit interface into eight regions as gaze target areas and segment the captured scene images to obtain the current facial image. We use the improved Swin Transformer model to learn the mapping from head posture and face images to the gaze direction of the flight cockpit interface. This method was used to estimate where the pilot directed their gaze during the flight, and the resulting heatmap was visually presented. The heatmap represented the interaction probabilities between the pilot and all areas within the cockpit scene, effectively highlighting the objects currently interacting with the pilot. We leveraged this heatmap to analyze the areas of particular interest to the pilot during flight tasks and to study their gaze patterns toward the target areas. This analysis forms the basis for subsequent investigations into the pilot's attention and operational intent during the flight process.

III. EXPERIMENTS AND METHODS

In the context of a simulated flight task, this paper uses a non-invasive appearance-based gaze tracking technique as the research methodology to enhance the efficiency of human-machine interaction for pilots and the accuracy of tracking their gaze. The objective is to investigate the pilot's gaze direction during the flight task and estimate their focus on specific areas within the flight cabin. As depicted in Figure 1, it illustrates the coordinate system of the pilot's head position and gaze direction. In this coordinate system, the parameters (p, y, r) correspondingly represent different orientations of the head, signifying the pitch angle, yaw angle, and roll angle, respectively. Furthermore, (x, y, z) indicates the position of the center of the face relative to the camera axis origin, and the point at which the pilot's gaze intersects the screen is referred to as the fixation point.

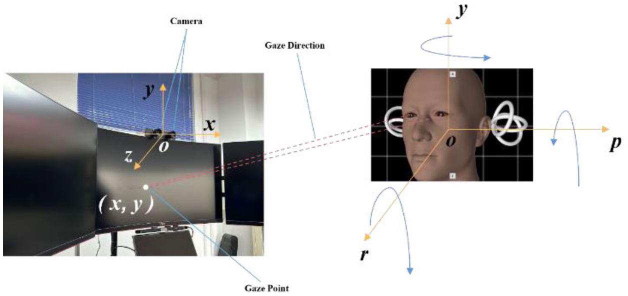


FIGURE 1. Gaze tracking system.

A. EXPERIMENTAL EQUIPMENT AND EXPERIMENTAL TASKS

1) EXPERIMENTAL EQUIPMENT

This article uses a six-axis simulated flight platform as the environment, as shown in Figure 2. The simulated flight platform is equipped with a three-screen spliced display, which can provide subjects with more realistic visual effects. In addition, this article adds a binocular camera and an illumination source to the simulated flight platform. The binocular camera is used to capture the subject's facial image during the flight task, and the lighting source is used to increase the visibility of the subject's pupils.

In-flight experiments, the cockpit instrument areas of primary interest to pilots are primarily centered around the central display. We have divided the instrument area of the flight cockpit into eight distinct gaze regions, primarily categorized as interior and exterior areas, as shown in Figure 3. The interior regions are numbered 1 to 5, corresponding to the landing gear, airspeed indicator, altimeter, attitude indicator, and heading indicator areas, respectively. The exterior regions are numbered 0, 6, and 7, corresponding to the Head-Up Display (HUD), left exterior, and right exterior areas.

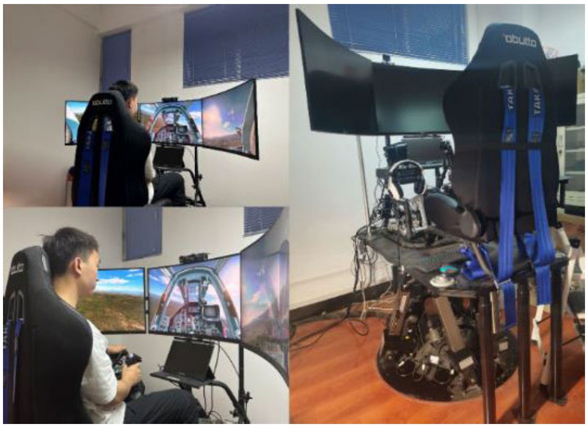


FIGURE 2. Six-axis model flight platform.

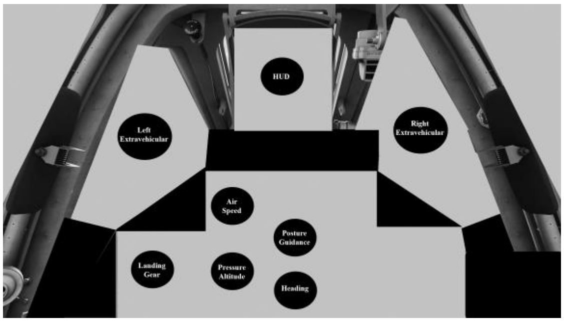


FIGURE 3. Flight cockpit annotation target area division.

TABLE 1. Flight mission requirements.

Mission phase	Air Route	Requirement
Take off	1	Take off to 600m cruise altitude, heading 095
Cruise	2	600m fixed height, turn to heading 005 cruise
	3	600m fixed height, turn to heading 275 cruise
	4	600m fixed height, turn to heading 185 cruise
Land	5	Turn to runway direction 095 and land

2) EXPERIMENTAL TASKS

The simulated flight experiment course is the descending flight training of the cruising aircraft on the five-sided route. The experimental route of the five-sided cruising take-off and landing is shown in Figure 4. The task defined four waypoints. The subjects started from point A of the runway and completed a five-sided flight around four waypoints and five routes. The five-sided mission requirements are shown in Table 1. The subjects were required to maintain the correct altitude, heading, and other flight tasks during the basic take-off, landing, and cruise phases.

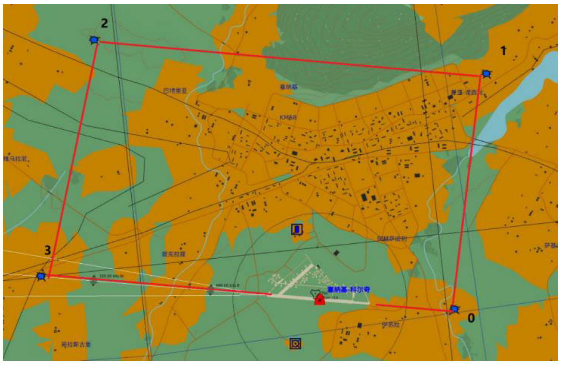


FIGURE 4. Five-sided cruise take-off and landing experimental route.

B. DATA COLLECTION AND PROCESSING

1) DATASETS

The data set was obtained by recruiting 10 flight cadets (4 females, 6 males, 24-26 years old) with healthy eyes in the laboratory. The recruited flight cadets, familiar with flight safety standards and procedures, have undergone guidance and training from professional pilots at the flight academy and have earned recognition from these professionals. Currently, the flight cadets possess proficient skills in flight take-off and landing tasks and can fulfill the requirements of the five-sided flight experiment. All flight attendants have agreed to participate in this research work. A complete five-side simulated flight experiment lasts about 40 minutes. The acquisition program synchronously collects the face images of the subjects after they gaze at the target area during the simulated flight. To avoid not detecting the complete face image and ensure the quality of the face image, the acquisition program will automatically discard the incomplete face image and the blurred face image by judging. The face image data set of the eight types of flight cockpit gaze areas is shown in Table 2. The total data set is 44353 images. To minimize the risk of potential data leakage during the training process, we implemented a strategy involving random splitting and identifier separation. Firstly, we thoroughly shuffled the acquired images and their corresponding labels to ensure the randomness of the data. Subsequently, while dividing the dataset, we assigned an identifier of 0 to undivided data and changed the identifier to 1 for the already split data. Using this strategy, we divided the entire dataset into two parts: 80% of the data as the training set, which consists of 35482 images, and the remaining 20% as the test set, comprising 8871 images. This process was strictly controlled to ensure no conflicting data between the training and test sets. The purpose of this data-splitting method is to ensure that the model can genuinely assess its performance during the testing phase, thereby making more accurate predictions for real-world applications. Through this approach, we can conduct model evaluation and predictions more reliably, reducing the risk of overly optimistic predictions caused by data leakage.

TABLE 2. DataSets composition.

Flight Cockpit	Target Area	Datasets	Training	Testing
Cabin Area	Landing Gear	5636	4508	1128
	Air Speed	5663	4530	1133
	Pressure Altitude	5508	4406	1102
	Posture Guidance	5449	4359	1090
	Heading	5723	4578	1145
Extravehicular Area	HUD	5468	4374	1094
	Left Extravehicular	5249	4199	1050
	Right Extravehicular	5657	4525	1132

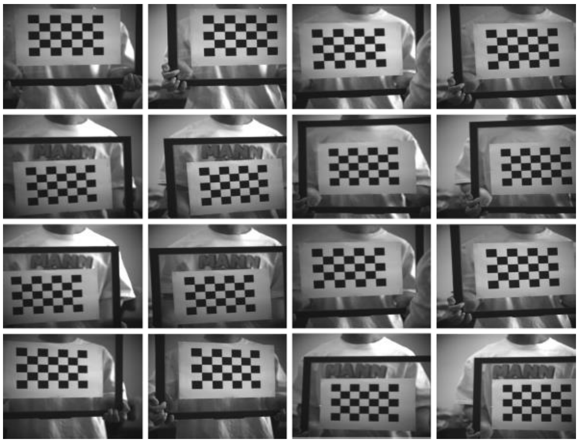


FIGURE 5. Checkerboard calibration diagram.

2) IMAGE PROCESSING

To avoid the distortion of the image, we calibrate the binocular camera, as shown in Figure 6. The distorted image captured by the binocular camera is corrected. We use the Zhengyou Zhang calibration method [37], using a two-dimensional planar calibration board composed of checkerboards. Multiple calibration board images of the binocular camera at different angles and different orientations are collected, and the internal and external parameters of the camera are calculated by the corner points of the checkerboard. The checkerboard images of different angles are shown in Figure 5. We use the checkerboard as a calibration board to handle complex three-dimensional environments. To obtain multi-dimensional coordinate information, we use different angles and different directions to capture chessboard images.

In binocular camera calibration, we use reprojection error as the evaluation criterion of calibration results. The re-projection error is the error pixel between the projected point (theoretical value) and the measurement point on the image, as shown in Figure 7. We control the re-projection error pixel within 0.15, select 23 valid calibration board images, and screen out the calibration board with a large re-projection error.

After calibrating the binocular camera, we obtain the internal and external parameters of the binocular camera.

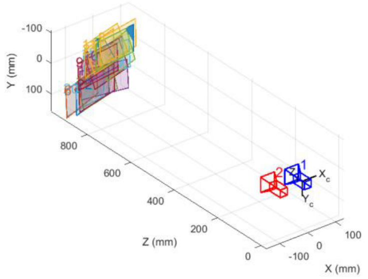


FIGURE 6. Binocular camera calibration.

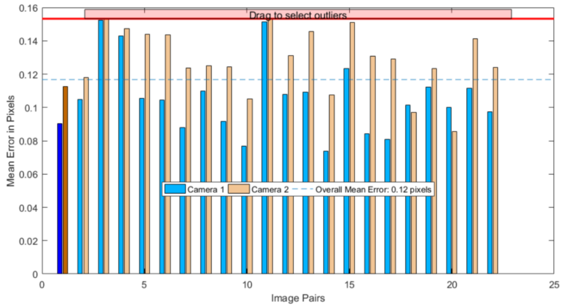


FIGURE 7. Reprojection error.

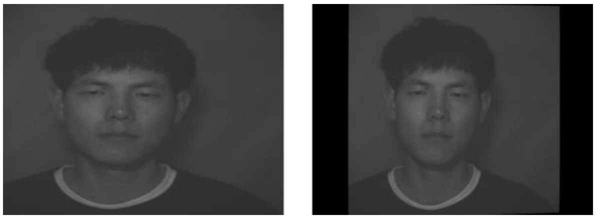

(a) Face image before correction
(b) Face image after correction

FIGURE 8. Comparison of face images before and after correction.

We correct the binocular camera images using the calculated camera rotation and offset matrices, in-camera parameters, and radial and tangential distortion parameters. The image before correction is shown in Figure 8 (a), and the image after correction is shown in Figure 8 (b).

Bulat and Tzimiropoulos [38] use a face alignment method based on deep learning to detect facial key points and effectively identify facial contours in different orientations and postures. The excellent performance of this method led us to choose to use the model method proposed by Bulat et al. to detect and recognize the corrected 2D face images and to segment and extract the facial image simultaneously. The facial image results after segmentation and extraction are shown in Figure 9.

3) HEAD POSE ESTIMATION

Head pose estimation is the process of determining the three-dimensional orientation or direction of a human head using computer vision techniques. Head pose data typically includes information about the position, orientation, and posture of the head. The size and categories of this data may vary depending on the specific methods used for data collection and processing, typically covering positional, directional, and postural data. We determine the head pose by analyzing the position and orientation of the face, primarily obtaining positional data for the head pose. This data comprises the pitch, yaw, and roll angles, representing the head's rotation relative to the horizontal plane in three-dimensional space, namely the coordinates of the head in 3D space. The categories of head pose data typically include actions such as turning the head, tilting it up, tilting it to the side, or lowering it, among others. Different head poses categories correspond to different head positions. In this paper, the head pose data we integrate primarily consists of three-dimensional positional data. When inputting head pose data into a neural network, it is converted into a tensor format, and the dimensions may vary based on the network architecture and task requirements.

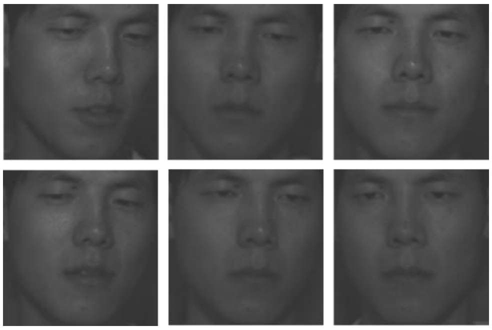


FIGURE 9. Facial figure.

To perform head pose estimation, the binocular camera first needs to be calibrated to obtain accurate camera parameters. Next, we can leverage the Dlib library, as Dlib is a popular open-source face recognition library and a toolkit for modern machine learning, computer vision, and image processing. This library provides a range of tools and models for face analysis, including face detectors and keypoint detectors. The face detector is used to locate the position of faces in the image, while the landmark detector is employed to detect facial features such as the eyes, nose, and mouth. By combining face detection and key point detection models, we calculate the positions and spatial relationships of key feature points on the face. This enables us to infer the rotation and tilt angles of the head, facilitating head pose estimation.

Accurate head pose estimation enhances the precision of gaze tracking and compensates for gaze offset. Camera calibration and the facial detection and key point detection capabilities of the Dlib library can provide relatively precise and stable results for head pose estimation. Therefore, we can accurately achieve head pose estimation and compensate for gaze offset. As shown in Figure 10, we identify specific facial landmark points and determine the orientation of different head poses. This allows us to calculate the pitch, yaw, and roll angles of the subject's head posture in the current frame.

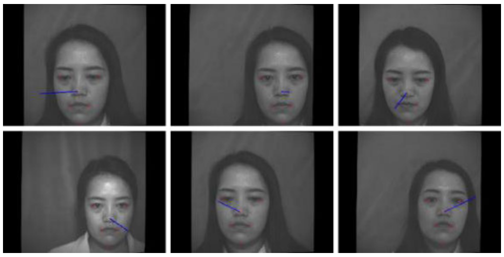


FIGURE 10. Head posture.

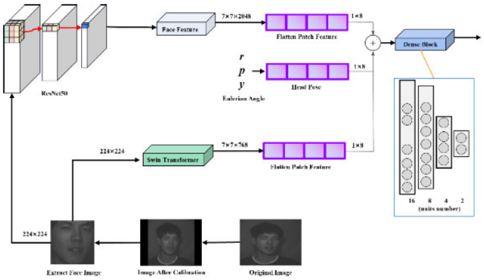


FIGURE 11. Hybrid model of gaze target region classification.

C. HYBRID MODEL

The gaze target region classification hybrid model proposed in this paper is shown in Figure 11. We first calibrate and crop the original image, use the face image with a size of 224 × 224 as the input of the model, and use the ResNet50 model to extract the features of the face image. Then we compare the classification results of the dataset in the pure Vision Transformer and the pure Swin Transformer models. The findings indicate that the Swin Transformer model performs better in classifying the gaze target area. Therefore, we combine the face image features extracted by Swin Transformer, the features extracted by ResNet50, and the head pose features. This combination is used for classifying and detecting the gaze target area, and we analyze the performance of the hybrid model.

1) RESNET

Traditional CNN stacks a series of convolutional and downsampling layers. However, as the network becomes deeper, problems like vanishing or exploding gradients and degradation arise. To address these issues, the ResNet model was introduced. It uses Batch Normalization (BN) layers to handle the vanishing or exploding gradient problem and proposes the residual structure to alleviate degradation. In this study, we adopt the ResNet50 architecture to extract facial image features and head pose features, as shown in Figure 12.

The ResNet50 model structure takes the original human face image as input. Firstly, the image size is changed to 112 × 112 by convolution operation with a convolution kernel of 7 × 7 and step size of 2. After each layer of convolution, the BN layer is connected to solve gradient disappearance or gradient explosion. The four Block blocks are residual structures with a different number of convolution kernels. The first layer of each residual structure block is a 1 × 1 convolution kernel, which is used to compress the channel dimension. The second layer is a 3 × 3 convolution kernel. The third layer is a 1 × 1 convolution kernel to restore the channel dimension. The size of the first layer convolution kernel on the main branch of the residual structure is the same as that of the second layer, and the number of convolution kernels in the third layer is four times that of the first layer.

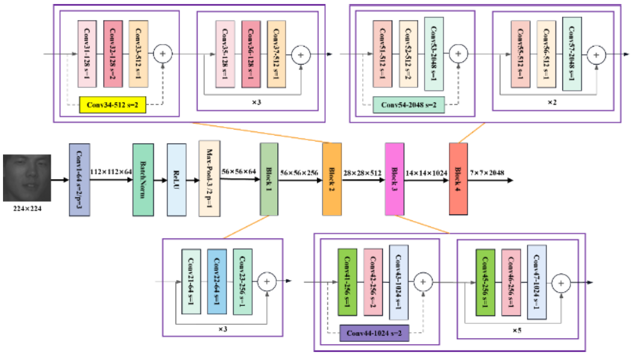


FIGURE 12. ResNet50 model structure diagram.

The residual structure can be divided into a dashed-line residual structure and a solid-line residual structure. In the dashed-line residual structure, there is a 1 × 1 convolutional layer in the shortcut branch, and the number of convolutional filters in this layer is the same as that in the third convolutional layer of the main branch. The dashed-line residual structure serves the purpose of adjusting the size of the input feature matrix. It achieves this by adding the output feature matrix from the shortcut branch to the output feature matrix from the main branch, ensuring that the input and output feature matrix sizes are the same. In the solid-line residual structure, the symbol "×N" indicates that the solid-line residual structure block is repeated N times.

2) VISION TRANSFORMER

Since the emergence of deep learning, CNN has been the mainstream model in the field of Computer Vision (CV) and has achieved good results. In contrast, the Transformer [39] model based on self-attention structure has shone in the field of Natural Language Processing (NLP). The Transformer model has a milestone significance in NLP but has been limited in its application in the computer vision domain. However, with further research, the Transformer architecture has started to be applied to computer vision tasks. In 2020, Dosovitskiy et al. [40] proposed a fully self-attention-based image classification model called Vision Transformer (VIT), which became a significant breakthrough in computer vision technology. The results demonstrate that the Transformer is indeed effective in the CV domain and yields impressive performance.

We have also conducted a comparative analysis of the performance of the hybrid VIT model, as shown in Figure 13. In this model, the feature map output from the ResNet50

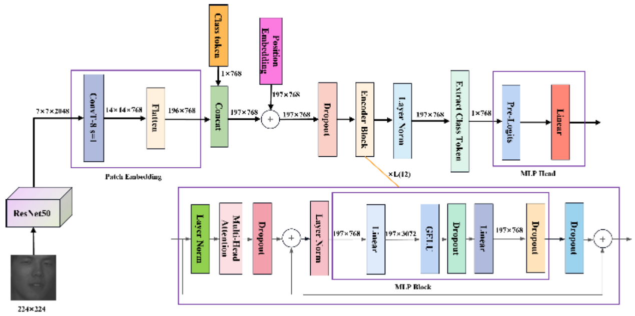


FIGURE 13. VIT-B/16 model diagram.

model is used as input, denoted by set $F = \{f_i\}_{i=1}^{N}$, where $f_i \epsilon R^{C\times H\times W}$ represents the i-th feature map, and C, H, W represents the number of channels, height, and width of the feature map. Integrating the VIT-B/16 model, as depicted in Figure 13, we employ the deconvolution function ConTranspose2d from the PatchEmbedding layer of the VIT-B/16 model to transform the image data. The kernel size is set to 8, and the stride is set to 1, resulting in the resizing of the image data to 14 × 14 × 768. Simultaneously, the height (H) and width (W) dimensions are flattened for further processing. Firstly, each image is divided into m blocks $x_i \epsilon R^{C\times P\times P}$ according to the specified size, where P represents the dimension of each block and $m = HW/p^2$. Then, the image block sequence is projected into a D-dimensional vector space through a learnable embedding matrix $E \epsilon R^{(P^2C)\times D}$ by linear mapping. Simultaneously, the position information of each block in the original feature map is re-encoded to obtain the position embedding $E_{pos} \epsilon R^{(m+1)\times D}$. It is then connected in series with a Class token x_{token} dedicated to classification, where x_{token} is a trainable parameter with a vector data format. x_{token} is stitched together with *tokens* generated in the image, where $tokens = \{x_i E\}_{i=1}^{m}$, together with Position Embedding, is embedded in E_{pos} to form the final embedded image block sequence z_0, that is:

$$z_0 = [x_{token}; x_1E; x_2E; \ldots; x_mE] + E_{pos} \tag{1}$$

The embedded image block sequence is sent to the VIT-B/16 encoder, as shown in Figure 13. and stacked L times repeatedly. Firstly, through the LayerNorm, each sample is normalized on the same feature dimension to improve the training speed and performance of the model. Then, further processing is performed using the Multiheaded Self-Attention (MSA) mechanism, which is represented as shown in Formula (2).

$$\begin{aligned} MSA\,(z) &= MSA\,(Q, K, V) \\ &= Concat(Head_1, Head_2, \ldots, Head_h)w^o \end{aligned} \tag{2}$$

where $Concat()$ denotes stacking in the form of column vectors, and $Head_i$ denotes:

$$Head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)i = 1, \ldots, h \tag{3}$$

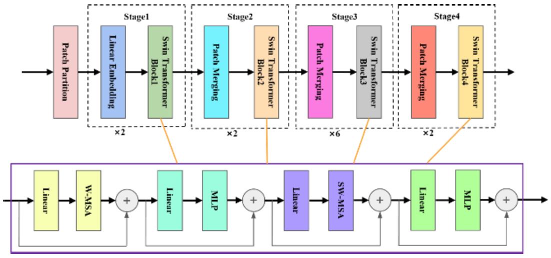


FIGURE 14. Swin Transformer model diagram.

where $W_i^K \epsilon R^{D\times d_k}$, $W_i^V \epsilon R^{D\times d_v}$, $W^o \epsilon R^{hd_v \times D}$, and $d_k = d_v = D/h$.

After obtaining the Q_i, K_i, V_i parameter corresponding to each $Head_i$, the *softmax* processing is performed for each $Head$, as shown in formula (4).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

Because the value after the multiplication of point QK^T is very large, the gradient becomes very small after passing through point *softmax*. To make the gradient more stable, it needs to be divided by $\sqrt{d_k}$ for scaling.

Secondly, the Multilayer Perceptron (MLP) adjusts the number of input nodes to four times using the first fully connected layer. Then, it restores the number of nodes using the second fully connected layer and combines the GELU activation function to expedite model convergence and mitigate the gradient disappearance problem during training. The MSA and MLP processes are shown in formula (5) and (6).

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \quad l = 1, \ldots, L \tag{5}$$

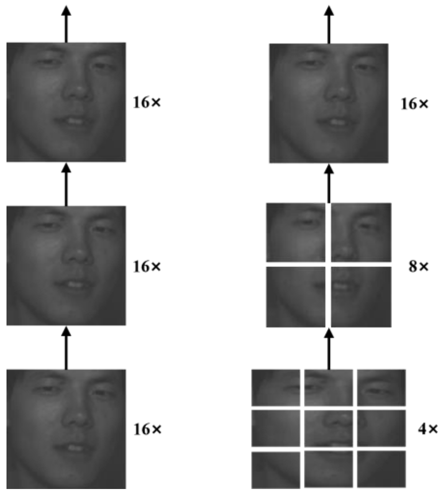
$$z_l = MLP(LN(z'_l)) + z'_l \quad l = 1, \ldots, L \tag{6}$$

Finally, through the MLP Head block, we get our final classification result of the gaze target area.

3) SWIN TRANSFORMER

The Swin Transformer [9] model is a deep learning architecture based on the Transformer model, which is mainly used in computer vision tasks. Swin Transformer has achieved state-of-the-art performance in various benchmarks, surpassing traditional CNN and previous Transformer-based models in many cases. By comparing the VIT-B/16 model, we find that the Swin Transformer model has a slightly better classification effect in the gaze target area. The comparison between VIT and Swin Transformer is shown in Figure 15. Similarly, our target gaze region classification hybrid model takes the feature map output by the ResNet50 model as input and combines the Swin Transformer model to obtain the classification results. The Swin Transformer model is shown in Figure 14.

The Swin Transformer model first inputs the feature map into the Patch Partition module for partitioning and then performs a linear transformation on the channel data of each pixel through the Linear embedding layer. Next, four Stages are used to construct feature maps of different sizes.



（a）VIT （b）Swin Transformer

FIGURE 15. Swin Transformer hierarchical feature map.

The × N in the Stage indicates that the structure block is repeated N times. Except that Stage 1 first passes through a Linear Embedding layer, the remaining three stages are first down-sampled through a Patch Merging layer. Subsequently, the Swin Transformer Block is stacked repeatedly, one using the Windows Multi-head Self-Attention (W-MSA) structure and the other using the Shifted Windows Multi-Head Self-Attention (SW-MSA) structure. MSA and W-MSA as shown in Formula (7) and (8).

$$\Omega\,(MSA) = 4hwC^2 + 2(hw)^2C \tag{7}$$

$$\Omega\,(W - MSA) = 4hwC^2 + 2M^2hwC \tag{8}$$

where h represents the height of the feature map, w represents the width of the feature map, C represents the depth of the feature map, and M represents the size of each window. The W-MSA module is to reduce the amount of calculation. Firstly, the feature map is divided into windows by $M \times M$ size, and then the local multi-head attention calculation is performed on each window separately. To solve the problem that there is no information transfer between windows and windows, and only self-attention calculation can be performed within each window, the Swin Transformer model introduces the SW-MSA module, that is, the W-MSA with windows offset. Offset through windows to solve the problem of information exchange between different windows. The W-MSA structure and the SW-MSA structure are used in pairs, first using a W-MSA structure and then using a SW-MSA structure.

The Swin Transformer model also uses relative position paranoia, as shown in Formula (9), and relative position paranoia is added based on Formula (4). Relevant research [9] has proved that the accuracy rate will be significantly improved after using relative positional paranoia.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} + B)V \tag{9}$$

where B is relative position paranoia.

TABLE 3. Comparison of classification accuracy of different models.

Model	Image Size	ImageNet top-1 acc.	ImageNet + HeadPos top-1 acc.
VIT-B/16[40]	224^2	77.05	—
VIT-L/16[40]	224^2	78.14	—
ResNet34[8]	224^2	84.86	86.02
ResNet50[8]	224^2	85.85	86.92
Swin-T[9]	224^2	85.90	87.60
Swin-S[9]	224^2	87.07	89.32
ResNet50+VIT-B/16	224^2	80.97	83.19
ResNet50+VIT-L/16	224^2	81.24	83.48
ResNet50+Swin-T	224^2	88.58	90.09
ResNet50+Swin-S	224^2	89.21	90.51

Finally, for the classification network, the model will be followed by a Layer Norm layer, a global pooling layer, and a fully connected layer to obtain the final output.

IV. RESULTS ANALYSIS AND DISCUSSION

A. ANALYSIS OF MODEL RESULTS

The classification effect of the target area is shown in Table 3. We train and compare different models. The network training epochs of each experiment are 100 epochs, the batch size of each training is 32, and the learning rate is variable. The evaluation index for the gaze target area classification model involves comparing the output value of the network model with the real label value. This analysis assesses the consistency between the area determined by the current model and the real label value.

To evaluate the performance of the attention-based region classification model based on the hybrid Swin Transformer, we compared and analyzed the detection results of the VIT model [40], the ResNet model [8], the Swin Transformer model [9], and the hybrid model. We used extracted face images as input to the models, with a size of 224 × 224. To minimize the risk of potential data leakage during the training process, we have employed a strategy involving random splitting and identifier separation. In this approach, 80% of the dataset is allocated for the training set, and 20% of the data is designated for the test set to facilitate the training process. Further details about the dataset have been provided in the dataset section. As shown in Table 3, the classification performance of the dataset in the VIT model and the hybrid VIT model was not as good as the ResNet model and Swin Transformer model. In the absence of head pose features, the ResNet model achieved an accuracy of 85%, with ResNet50 slightly outperforming ResNet34. After combining the features of head pose, both ResNet34 and ResNet50 show noticeable improvements, but ResNet50 still maintains a higher classification accuracy than ResNet34. Therefore, we have chosen ResNet50 as a part of the hybrid model. The Swin Transformer model achieved an accuracy of 87% when using only facial features, which improved to 89% after incorporating head pose features. In the hybrid models, the performance of the ResNet mixed with the VIT model showed some improvement compared to VIT alone, but the

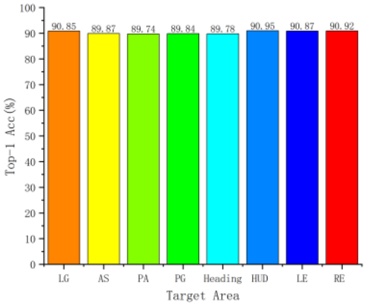


FIGURE 16. Target area classification results.

overall performance was not very good. On the other hand, in the hybrid Swin Transformer model, there was a significant improvement in overall performance compared to both the pure Swin Transformer and pure ResNet models. Furthermore, after fusing head pose features, the classification accuracy reached 90%. Overall, the hybrid Swin Transformer model demonstrated the best performance among the evaluated models in this study. This was particularly evident when considering the incorporation of head pose features, significantly enhancing the classification accuracy.

We compared and analyzed the classification performance of various target regions as shown in Figure 16. The horizontal axis represents eight cockpit attention target regions: Landing Gear (LG), Air Speed (AS), Pressure Altitude (PA), Posture Guidance (PG), Heading, Head-Up Display (HUD), Left Extravehicular (LE), and Right Extravehicular (RE). The vertical axis represents the classification accuracy. We observed that the classification accuracy of the LG, HUD, LE, and RE regions was above 90%, indicating excellent performance in these areas. On the other hand, the AS, PA, PG, and Heading regions achieved classification accuracies below 90%. Through our analysis, we deduced that this might be due to the proximity of the AS, PA, PG, and Heading regions, leading to potential errors in the model's judgments and resulting in a decrease in accuracy. However, overall, the classification performance still reached satisfactory results. In conclusion, the model achieved desirable results, with high accuracy in several target regions, while some regions with close similarity showed slightly lower accuracy due to potential confusion during classification. Nevertheless, the overall classification performance met the desired criteria.

B. FLIGHT VISUALIZATION ANALYSIS

We analyze the gaze distribution of flight cadets performing flight missions during flight through heat maps. We import the flight cadets ' gaze distribution data during some mission periods and classify them according to the number of gazes in the target area. The red area indicates the area with the highest number of gazes, the orange area indicates the area with relatively few gazes, the green area indicates the area with the least number of gazes, and the colorless area indicates that the area has not gazed at the current stage. Through visual analysis, it can be judged whether the pilot's current operation is correct, and the pilot's operation intention and attention can be judged later.

In the simulation of the flight experiment, we primarily demonstrated the flight trainees' focus on instruments during takeoff tasks, altitude adjustment tasks, and turning tasks.

1) THE TAKEOFF TASK

In the takeoff tasks, assuming the flight student has taxied the aircraft to the takeoff line aligned it with the runway, and has checked that all instruments and warning lights are functioning normally and meet the flight requirements. At this point, the air traffic control tower has issued the command, allowing the flight student to execute the takeoff task. During this process, the flight student follows instructions to complete the takeoff task, which includes starting the aircraft's taxi, gaining sufficient speed for takeoff, and the process of retracting the landing gear.

We randomly selected a group of flight students and observed their gaze patterns during a task, as shown in Figure 17. The flight student's primary areas of focus were the airspeed indicator, landing gear status, and altitude indicator. The head-up display (HUD) and heading indicator received relatively less attention, while other areas that were glanced at may have resulted from the flight student's scanning. In the takeoff task, the instructions were to set the aircraft's heading to 095. The student was told to initiate takeoff at an airspeed of approximately 160-180 kilometers per hour and to retract the landing gear when the aircraft's altitude reached around 10 meters, completing the takeoff task. The program records the duration and number of times the student looks at each instrument area during the takeoff task. The results show that the student's total gaze duration is 56995ms, and the total gaze count is 4676 times. The gaze duration and gaze count of the main gaze areas are shown in Table 4. Through result analysis, we can check if the flight student uses the airspeed indicator to assess if the current aircraft meets takeoff criteria during the flight task. The student also relies on the heading indicator to confirm that the aircraft maintains the prescribed heading and uses the altitude indicator to ensure the conditions for landing gear retraction are met. Furthermore, the landing gear status can be combined with these assessments to analyze the current condition of the landing gear, ensuring the safety of the flight.

2) ADJUST THE HEIGHT TASK

In the climb altitude task, we randomly selected a group of flight students and observed their gaze behavior during the task, as shown in Figure 18. The primary area of focus for the flight students was the altimeter, while they paid relatively less attention to the Head-Up Display (HUD), airspeed indicator, attitude indicator, and heading indicator. Other areas or background areas they glanced at were likely the result of the flight students' scanning behavior. In the climb altitude task, the flight student was instructed to increase the current flight altitude from 600 meters to 900 meters. We recorded the

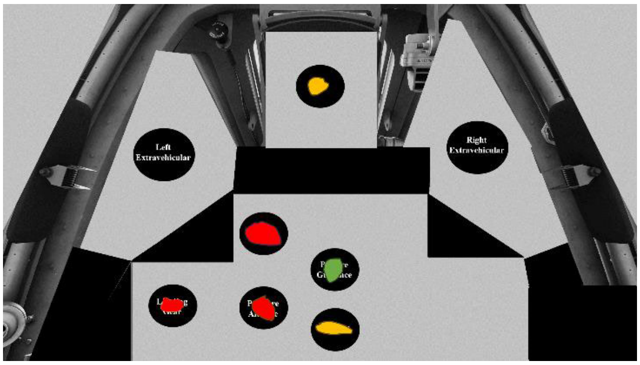


FIGURE 17. Takeoff task fixation distribution.

TABLE 4. Statistical results of gaze duration and gaze frequency in takeoff tasks.

Flight Cockpit	Target Area	Gaze Duration(ms)	Gaze Frequency
Cabin Area	Landing Gear	5128	421
	Air Speed	25649	2108
	Pressure Altitude	9625	786
	Posture Guidance	1039	86
	Heading	11458	939
Extravehicular Area	HUD	3521	289
	Left Extravehicular	164	14
	Right Extravehicular	176	14
Background Areas		235	19

student's gaze duration and frequency in various instrument areas from the moment they received the task until task completion. The results revealed that the total gaze duration for this student was 62268 milliseconds, with a total gaze frequency of 5099 times, as depicted in Table 5. The longer gaze duration and higher gaze frequency indicate that this flight student preferred to focus on specific areas during the flight task rather than continuously monitoring all spatial sectors. Through result analysis, we can confirm that in the climb altitude task, the flight student looks at the altimeter and HUD area to determine if the aircraft has reached the required standard altitude. They also use the heading and attitude guidance table to assess whether the current aircraft deviates from the course or is in a horizontal state. Additionally, they check the airspeed indicator to ensure the aircraft is within the normal flying speed and avoid stalling.

TABLE 5. Statistical results of gaze duration and gaze frequency in height tasks.

Flight Cockpit	Target Area	Gaze Duration(ms)	Gaze Frequency
Cabin Area	Landing Gear	158	12
	Air Speed	8696	713
	Pressure Altitude	28515	2337
	Posture Guidance	7386	607
	Heading	7692	632
Extravehicular Area	HUD	9018	738
	Left Extravehicular	213	17
	Right Extravehicular	209	17
Background Areas		381	26

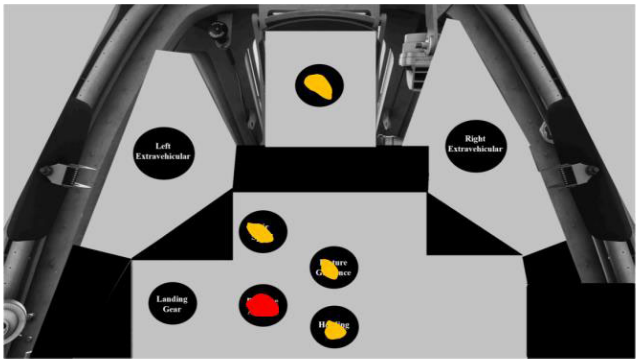


FIGURE 18. Adjusting the height task fixation distribution.

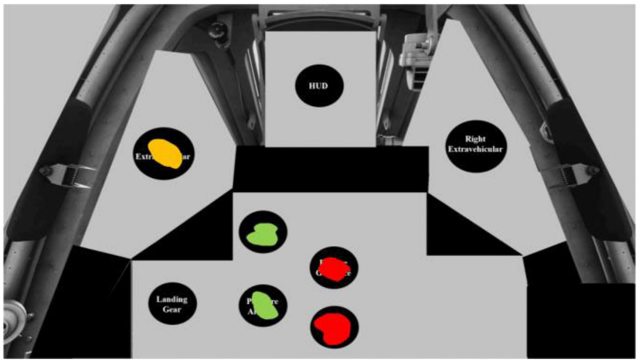


FIGURE 19. Gaze distribution of left turn task.

3) TURNING TASK

In the left turn flight task, as depicted in Figure 19, the flight student primarily focused on the attitude indicator and the heading indicator. The next area of focus was the left-side cabin area. The altimeter and airspeed indicator received fewer gazes, while other areas or background areas that received attention might be attributed to the flight student's scanning behavior. During the turning task, the flight student was instructed to change the current heading from 005 to 275. We recorded the student's gaze duration and frequency in various instrument areas from when they received the task until completion. The results indicated that the total gaze duration for this student was 119612 milliseconds, with a total gaze frequency of 9816 times, as illustrated in Table 6. We can judge that in the task of turning left, the flying cadets judge whether the aircraft reaches the target course through heading and attitude guidance. The reason for watching the left extravehicular area is that the flying cadets will subconsciously watch the left extravehicular scene when turning left. Because altitude and airspeed are constantly changing during turning missions, flight cadets need to pay attention to the current altitude and airspeed of the aircraft to avoid stalling or crashing into objects.

TABLE 6. Statistical results of gaze duration and gaze frequency in left turn tasks.

Flight Cockpit	Target Area	Gaze Duration(ms)	Gaze Frequency
Cabin Area	Landing Gear	116	10
	Air Speed	4486	368
	Pressure Altitude	4246	348
	Posture Guidance	50256	4128
	Heading	52164	4277
Extravehicular Area	HUD	556	46
	Left Extravehicular	7254	595
	Right Extravehicular	107	9
Background Areas		427	35

C. DISCUSSION

The mixed Swin Transformer model is slightly superior to the pure Transformer model based on both the model results and visual analysis. Furthermore, the overall classification performance improves when incorporating head pose features compared to solely extracting face features. However, the classification effect of the model will decrease when the target area is similar. For such problems, we will further optimize the network model and improve the accuracy of the classification effect. In different flight missions, we can see that the attention of the flight cadets to the cockpit instrument area is different. In the follow-up work, we can further analyze the attention state or operation intention of the flight cadets through the attention of the flight cadets.

V. CONCLUSION

The Swin Transformer model has been a crucial player in computer vision, exhibiting outstanding performance in target detection and instance segmentation tasks. Therefore, we use the hybrid structure of the Swin Transformer model for classification purposes. The ResNet50 model extracts features from face images, fuse head pose features, and combines with the Swin Transformer model to classify the gaze target area. To verify the effectiveness of our proposed hybrid model structure, we conducted a comparative analysis of the classification outcomes of our datasets using VIT, Swin Transformer, ResNet50 + VIT, and ResNet50 + Swin Transformer models. Our results show that the classification efficacy of the dataset under the hybrid model structure reached 90%, which is of great significance. At the same time, we analyzed the current flight cadets' fixation in the flight scene through the heat map, which paved the way for the follow-up pilots' attention evaluation. In future work, we will optimize the network model to enhance accuracy while considering real-time requirements. Additionally, we will analyze the pilot's attention and operational intention during flight based on these improvements.

ACKNOWLEDGMENT

The authors would like to thank the creator of the Dlib library, Davis E. King, and his team, as well as all developers who have contributed to the library. The open-source nature and powerful functionality of Dlib have been crucial in supporting their research. In this study, Dlib's machine learning and computer vision tools have offered a dependable foundation, significantly easing our exploration of head pose estimation through facial and key point detection methods. They also like to thank the Flight Academy for their support, providing indispensable assistance to their team in our aviation-related work.

REFERENCES

[1] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 3943–3968, May 2022.

[2] L. Jiao, D. Wang, Y. Bai, P. Chen, and F. Liu, "Deep learning in visual tracking: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5497–5516, Sep. 2023.

[3] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 697–702.

[4] A. A. Akinyelu and P. Blignaut, "Convolutional neural network-based methods for eye gaze estimation: A survey," *IEEE Access*, vol. 8, pp. 142581–142605, 2020.

[5] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[6] P. Chakraborty, S. Ahmed, M. A. Yousuf, A. Azad, S. A. Alyami, and M. A. Moni, "A human–robot interaction system calculating visual focus of human's attention level," *IEEE Access*, vol. 9, pp. 93409–93421, 2021.

[7] J. Jiang, X. Zhou, S. Chan, and S. Chen, "Appearance-based gaze tracking: A brief review," in *Proc. 12th Int. Conf. Intell. Robot. Appl.* Shenyang, China: Springer, Aug. 2019, pp. 629–640.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[10] A. Tsukada, M. Shino, M. Devyver, and T. Kanade, "Illumination-free gaze estimation method for first-person vision wearable device," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2084–2091.

[11] Y. Matsumoto and A. Zelinsky, "Real-time face tracking system for human–robot interaction," in *Proc. IEEE Int. Conf. Systems, Man, Cybern. (SMC)*, vol. 2, Oct. 1999, pp. 830–835.

[12] V. S. Vasisht, S. Joshi, Shashidhar, Shreedhar, and C. Gururaj, "Human computer interaction based eye controlled mouse," in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Jun. 2019, pp. 362–367.

[13] K. Harezlak and P. Kasprowski, "Application of eye tracking in medicine: A survey, research issues and challenges," *Comput. Med. Imag. Graph.*, vol. 65, pp. 176–190, Apr. 2018.

[14] D. V. J. Shree, L. R. D. Murthy, K. S. Saluja, and P. Biswas, "Operating different displays in military fast jets using eye gaze tracker," *J. Aviation Technol. Eng.*, vol. 8, no. 1, p. 31, Dec. 2018.

[15] S. Peißl, C. D. Wickens, and R. Baruah, "Eye-tracking measures in aviation: A selective literature review," *Int. J. Aerosp. Psychol.*, vol. 28, nos. 3–4, pp. 98–112, Oct. 2018.

[16] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 3, pp. 254–265, Sep. 2018.

[17] M. Q. Khan and S. Lee, "Gaze and eye tracking: Techniques and applications in ADAS," *Sensors*, vol. 19, no. 24, p. 5540, Dec. 2019.

[18] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May 2016.

[19] V. Clay, P. König, and S. Koenig, "Eye tracking in virtual reality," *J. Eye Movement Res.*, vol. 12, no. 1, pp. 1–18, 2019.

[20] K. Qian, T. Arichi, A. Price, S. Dall'Orso, J. Eden, Y. Noh, K. Rhode, E. Burdet, M. Neil, A. D. Edwards, and J. V. Hajnal, "An eye tracking based virtual reality system for use inside magnetic resonance imaging systems," *Sci. Rep.*, vol. 11, no. 1, Aug. 2021, Art. no. 16301.

[21] S. Jia, D. H. Koh, A. Seccia, P. Antonenko, R. Lamb, A. Keil, M. Schneps, and M. Pomplun, "Biometric recognition through eye movements using a recurrent neural network," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Nov. 2018, pp. 57–64.

[22] J. Liu, J. Chi, H. Yang, and X. Yin, "In the eye of the beholder: A survey of gaze tracking techniques," *Pattern Recognit.*, vol. 132, Dec. 2022, Art. no. 108944.

[23] Y. Wang, T. Shen, G. Yuan, J. Bian, and X. Fu, "Appearance-based gaze estimation using deep features and random forest regression," *Knowl.-Based Syst.*, vol. 110, pp. 293–301, Oct. 2016.

[24] R. Naqvi, M. Arsalan, G. Batchuluun, H. Yoon, and K. Park, "Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor," *Sensors*, vol. 18, no. 2, p. 456, Feb. 2018.

[25] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Ophey, V. L. Flanagin, P. Z. Eulenburg, and S.-A. Ahmadi, "DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning," *J. Neurosci. Methods*, vol. 324, Aug. 2019, Art. no. 108307.

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

[27] S. Sayeed, F. Sultana, P. Chakraborty, and M. A. Yousuf, "Assessment of eyeball movement and head movement detection based on reading," in *Recent Trends in Signal and Image Processing*. Singapore: Springer, 2021, pp. 95–103.

[28] L. Dai, J. Liu, and Z. Ju, "Binocular feature fusion and spatial attention mechanism based gaze tracking," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 2, pp. 302–311, Apr. 2022.

[29] D. Cazzato, M. Leo, C. Distante, and H. Voos, "When I look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking," *Sensors*, vol. 20, no. 13, p. 3739, Jul. 2020.

[30] A. George and A. Routray, "Real-time eye gaze direction classification using convolutional neural network," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*, Jun. 2016, pp. 1–5.

[31] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 383–398.

[32] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[33] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba, "Following gaze in video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1435–1443.

[34] Z. Hu, C. Lv, P. Hang, C. Huang, and Y. Xing, "Data-driven estimation of driver attention using calibration-free eye gaze and scene features," *IEEE Trans. Ind. Electron.*, vol. 69, no. 2, pp. 1800–1808, Feb. 2022.

[35] Z. Hu, D. Yang, S. Cheng, L. Zhou, S. Wu, and J. Liu, "We know where they are looking at from the RGB-D camera: Gaze following in 3D," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[36] Z. Hu, K. Zhao, B. Zhou, H. Guo, S. Wu, Y. Yang, and J. Liu, "Gaze target estimation inspired by interactive attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8524–8536, Dec. 2022.

[37] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[38] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[39] A. V aswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 5998–6008.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

GONGPU WU received the B.S. degree in software engineering from Xi'an Technological University, in 2021, where he is currently pursuing the Ph.D. degree in photoelectric engineering. His research interests include artificial intelligence, computer vision, and human–machine hybrid direction.

CHANGYUAN WANG received the master's degree in applied mathematics from the Northwest University of Technology, in April 1988, and the Ph.D. degree in mechanical engineering from the Xi'an University of Technology, in 2011. He is currently a Professor and the Ph.D. Supervisor with Xi'an Technological University.

LINA GAO received the B.S. and M.S. degrees in software engineering from Qinghai Normal University, China, in 2019 and 2021, respectively. She is currently pursuing the Ph.D. degree in optical engineering with Xi'an Technological University. Her research interests include machine learning, artificial intelligence, and computer vision.

JINNA XUE received the B.S. degree in software engineering from Xi'an Technological University, in 2021, where she is currently pursuing the M.Eng. degree with the School of Computer Science and Engineering. Her research interests include artificial intelligence and software engineering.