Dean&Francis
ISSN 2959-6157

# An Enhanced OpenPose-Based Methodology for Detecting Pilot Fatigue During Flight Operations

**Yuke Chen**[1],

**Xiaoqi Yuan**[2,*],

**Ziquan Zhu**[3]

[1]Water conservancy & Hydropower engineering, Hohai University, Nanjing, 210024, China
[2]School of Mechatronics Engineering, Anhui University of Science and Technology, Huainan, 232000, China
[3]Shanghai World Foreign Language Academy, Xuhuai, 200030, China
*Corresponding author: 2023303560@aust.edu.cn

**Abstract:**

Single-Pilot Operations (SPO) represent a focal point in the current development direction of commercial aircraft, where monitoring the workload of a single pilot is crucial for ensuring flight safety. The detection results must exhibit both accuracy and real-time performance. This study leverages the widely adopted OpenPose model, which is employed for posture detection due to its capability to effectively extract human skeletal keypoints. However, the conventional OpenPose model suffers from issues such as large data volume, high computational demands, and significant redundancy. This paper systematically reviews relevant improvement methods across three key aspects of the OpenPose model: image preprocessing, feature extraction, and pose estimation. Additionally, it summarizes a fatigue assessment method based on this model. Research findings indicate that the enhanced OpenPose model, compared to the traditional version, demonstrates greater adaptability to complex environments, faster response and data processing capabilities, and improved accuracy in results. By incorporating fatigue detection criteria, the model enables real-time alerting.

**Keywords:** OpenPose model; single-pilot driving mode; fatigued driving

## 1. Introduction

Currently, the primary operational model employed in civil transport aircraft is the two-pilot crew system. While a multi-crew pilot configuration can alleviate the workload of individual pilots, it presents several drawbacks, such as directly increasing pilot training costs and imposing higher requirements for cockpit equipment. More critically, inadequate coordination among pilots in a multi-crew environment may lead to severe safety incidents. Therefore, the Single-Pilot Operations (SPO) model represents a core technological direction for the next generation of commercial aircraft. It is critically important to assess pilot workload under SPO conditions, ensuring that the operational pressure and burden on the pilot do not exceed those experienced in dual-crew operations, thereby maintaining flight safety [1]. Consequently,

ISSN 2959-6157

an intelligent posture detection system capable of dynamic monitoring, rapid analysis, and real-time alert notification is required to provide immediate feedback on pilots' operational status.

Currently, algorithms that combine deep learning and machine learning for human pose detection have been widely applied in many scenarios, and their accuracy can also be guaranteed. Among these, the OpenPose architecture is employed for pose detection due to its demonstrated efficacy in accurately extracting human skeletal key points [2]. Using this model to extract features of pilots, thereby detecting fatigued driving. The detection of pilot operations necessitates a high degree of precision and real-time performance to promptly identify potential hazards. However, the traditional OpenPose model suffers from issues such as extensive data and computational requirements, as well as high redundancy.

The OpenPose model is primarily structured into three key components: image preprocessing, feature extraction, and pose estimation. This paper will implement improvements across these three dimensions to enable accurate and real-time feedback of pilot physiological data during flight operations. Additionally, it introduces a methodology for detecting pilot fatigue based on the OpenPose model.

## 2. Overview of the Openpose Model

The OpenPose model, a bottom-up real-time multi-person 2D pose estimation algorithm, was developed through neural networks and supervised learning by a research team at Carnegie Mellon University. It enables comprehensive estimation of human body movements, facial expressions, hand gestures, and other skeletal dynamics. The core methodology involves utilizing Convolutional Neural Networks (CNN) for feature extraction, combined with Part Affinity Fields (PAF) to predict the relationships between human body parts and their corresponding limbs.

Traditional OpenPose models employ a VGG-19 convolutional neural network for preprocessing input images to generate feature maps F. These feature maps are subsequently fed into two computational branches, $\rho_t$ and $\varphi_t$. The branch $\rho_t$ is dedicated to predicting Part Confidence Maps (PCM), while the other branch $\varphi_t$ is utilized for estimating Part Affinity Fields (PAF). The subscript t denotes the stage index, representing sequential modules in the pipeline, with typical values set to 5, 6, or 7.

The output results of the Part Confidence Maps (PCM) and Part Affinity Fields (PAF) for skeletal key points at each stage are denoted as St and Lt respectively. In the first stage, the input is the feature map F output by the VGG-19 network. In the other stages, the input data consist of the feature map F output by VGG-19 and the confidence map PCM and affinity field PAF of the skeletal points output by the previous stage.

The St and Lt expressions are shown in Equation (1) [2].

$$\begin{cases} S^t = \rho^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2 \\ L^t = \phi^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2 \end{cases} \tag{1}$$

## 3. Image Preprocessing

To better extract features, this study proposes methods to address varying light intensities, noise, and facial image synthesis by adding infrared cameras, constructing a StyleGAN2-based generator network, applying sharpening, and using a dual-path generative adversarial network (GAN), tailored to the cockpit environment.

### 3.1 Improved Handling of Different Lighting Conditions

To address the complex lighting conditions within the cockpit and accurately capture facial images of pilots from multiple angles, an integrated cockpit light sensor can be employed in conjunction with infrared cameras equipped with automatically adjustable infrared LED illumination sources.In low light, the aperture automatically widens to increase light intake; in darkness, the infrared LED supplements light to facilitate image capture across conditions (see Figure 1). For clearer dark-environment images, infrared-to-visible light conversion is performed using a dual-contrast learning framework. Integrating a StyleGAN2 generator into this framework extracts finer-grained features, and its loss function enhances image details, producing high-quality color images in darkness. Alternatively, illumination compensation balances brightness/darkness, while HSV color modeling localizes facial regions and morphological processing fills holes in segmented, incomplete faces. Figure 2 illustrates the workflow for acquiring pilot facial images under varying lighting [3].

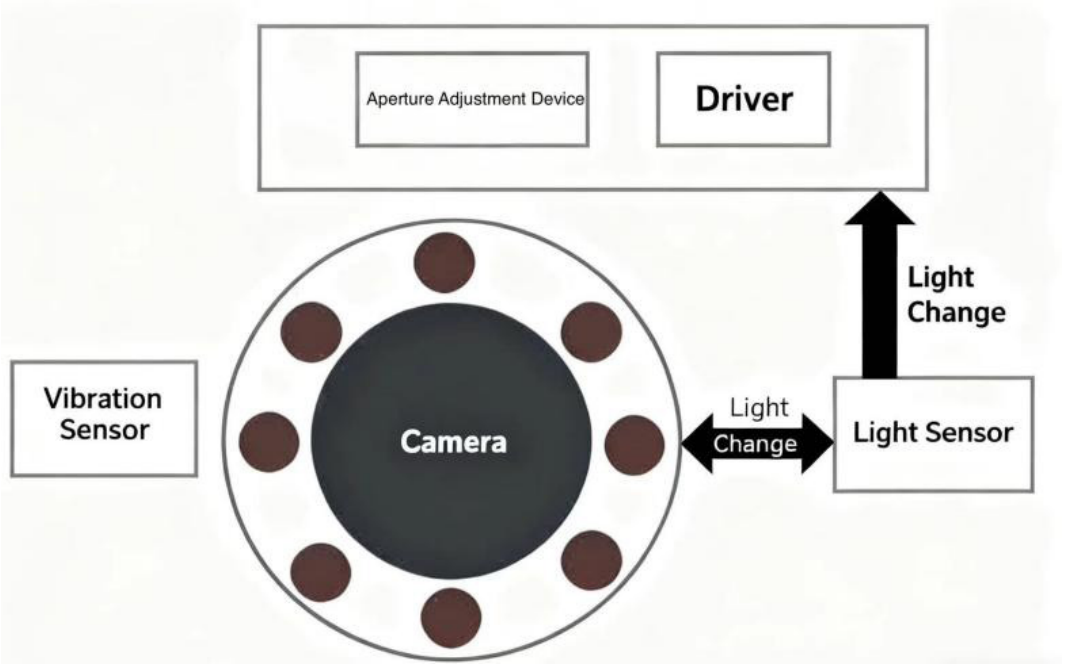


**Fig.1 Hardware structure diagram of the image acquisition device**

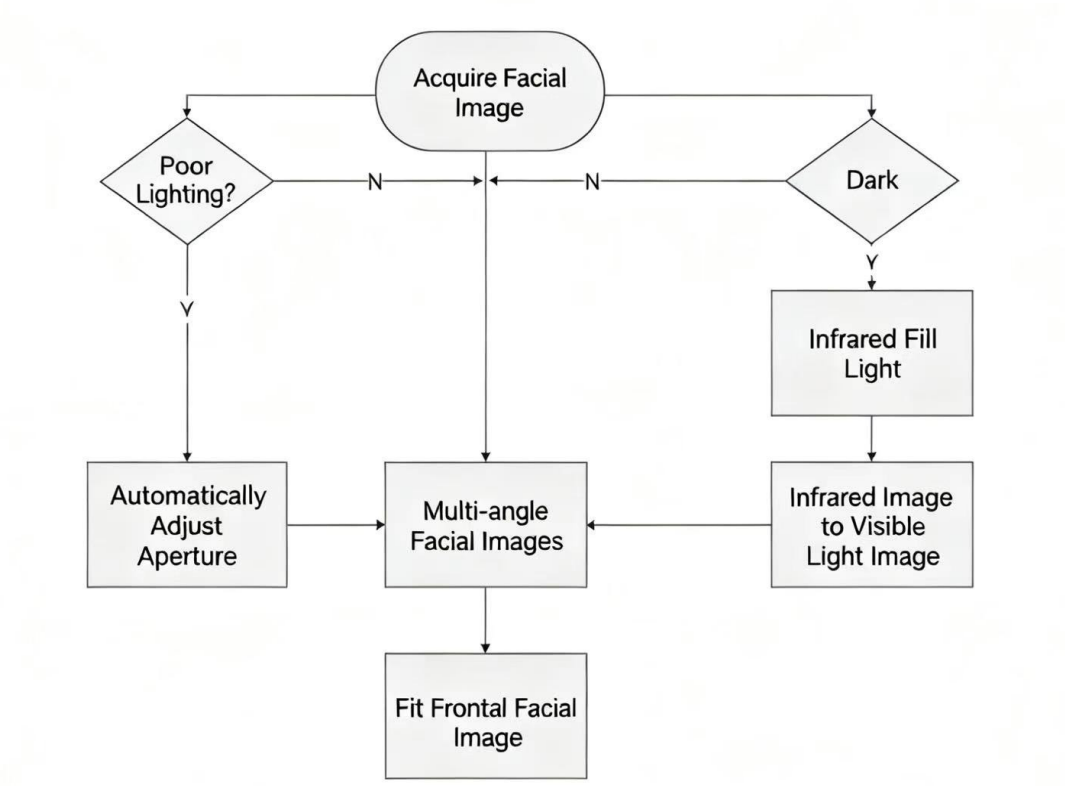


**Fig.2 Process diagram for obtaining the pilot's frontal image**

### 3.2 Noise Reduction Improvements

To enhance clarity, median filtering sorts pixels by RGB values, and sharpening strengthens edges. Median filtering eliminates impulse noise and outliers in the cockpit, while Laplacian sharpening partially restores and predicts images/residuals. Testing shows strong robustness against

ISSN 2959-6157

noise interference [4].

### 3.3 Facial Image Synthesis Improvements

A deep learning-based dual-path GAN synthesizes realistic frontal facial images from single inputs, supporting multi-angle rotation. It includes: (1) a local path learning eye, mouth, and nose features; (2) a global path feeding images into a CNN. The two paths are fused to generate frontal pilot faces [3]. Facial key points are expanded from OpenPose's original set to 68, with emphasis on eyes and mouth corners for detecting eye closure and yawning frequency [5]. Figure 3 shows 68 facial key points, and Figure 4 displays a 3D model derived from human key regions for pilot action judgment.

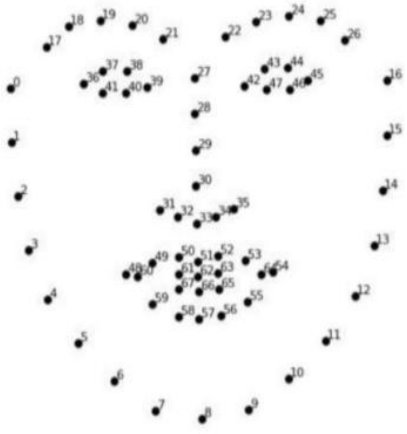

**Fig.3 Facial Recognition Feature Map (68 Facial Key Points)**

## 4. Feature Extraction

HRNet18 maintains a two-way branching structure for the last two layers in the same stage, and the rest of the layers are merged into a single-way structure, which is able to realize the sharing of parameters between the skeletal point confidence map PCM and the skeletal point affinity domain PAF. The original 7×7 convolution kernel is improved by changing it to a combination of 1×1, 3×3, and 3×3 structural convolution kernels, and the expansion coefficient of the last 3×3 convolution kernel is set to 2, which can maintain a consistent initial sensory field of view. The runtime performance and detection accuracy during feature extraction are improved by lightweighting the OpenPose model.

### 4.1 The Core Features Of HRNet-18

HRNet-18 parallel multi-resolution sub-network shown in Fig. 4, in which the resolution is usually reduced by down-sampling and then restored by up-sampling in conventional networks, maintains a stream of features of high, medium and low multiple resolutions in parallel throughout the network, which contains 4 stages, with progressively more parallel branches at each stage (expanding from 1 branch to 4 branches). It operates at a higher resolution, allowing information to be exchanged across layers and generating more accurate feature maps [6].

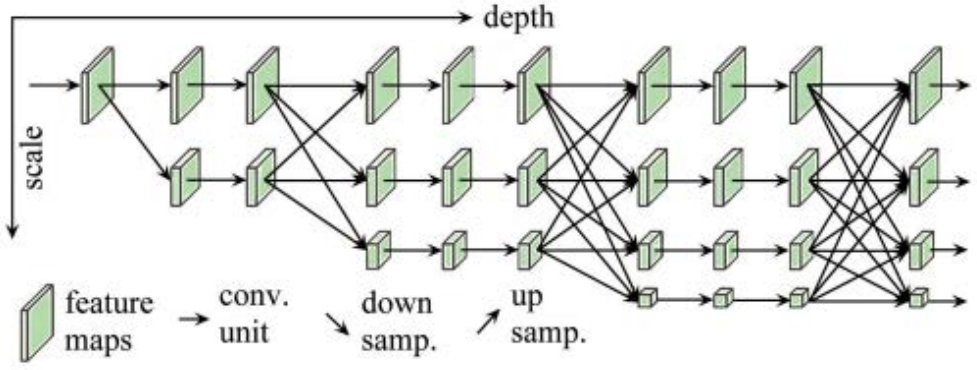


**Fig.4 Network structure of HRNet**

Information between different branches is exchanged through repetitive Multi-Scale Fusion: fusion methods include summing or splicing HRNet-18 has a parameter count of about 9.6 million (9.6M), with low computational loads (FLOPs), making it suitable for real-time applications (e.g., mobile or edge devices). In human posture estimation, COCO key point detection is used to accurately obtain the driver key point information, and high-resolution features are used to accurately localize the position of the five senses. The experimental data show that the computation amount of HRNet-18 is only 16% of VGG19 (Table 1). Therefore, HRNet-18 can be used to design a lightweight OpenPose network structure, and compared with the original inference performance, the lightweight OpenPose improves nearly 3 times. The following table shows the FLOPS and GPU inference performance of dif-

ferent versions of OpenPose [6].

**Table 1. Performance and Accuracy Testing of Openpose**

| Algorithmic models | FLOPS | TensorRT FP32 Reasoning speed | Cocodata Test accuracy |
|---|---|---|---|
| VGG19 Openpose | 236.1G | 43ms | 61.8AP |
| MobileNet | 27.7G | 12ms | 42.8P |
| HRNet18(OURS) | 37.4G | 14ms | 56.4G |

### 4.2 Behavior Detection Method Based On Lightweight OpenPose Spatiotemporal Graph Network

When obtaining the features of the data input to the lightweight OpenPose network shown in Fig. 5, the original 7×7 convolutional kernel is modified to a combination of 1×1, 3×3, and 3×3 kernels, and the expansion coefficient of the last 3×3 convolutional kernel is set to 2, which keeps the initial perception of the field of view consistent. After this improvement, the number of parameters is reduced to 15% of the second-order OpenPose algorithm, but the performance and accuracy are not affected, and the algorithm's running speed is greatly improved. The improved part of the lightweight OpenPose algorithm is shown in Figure 5 [6].

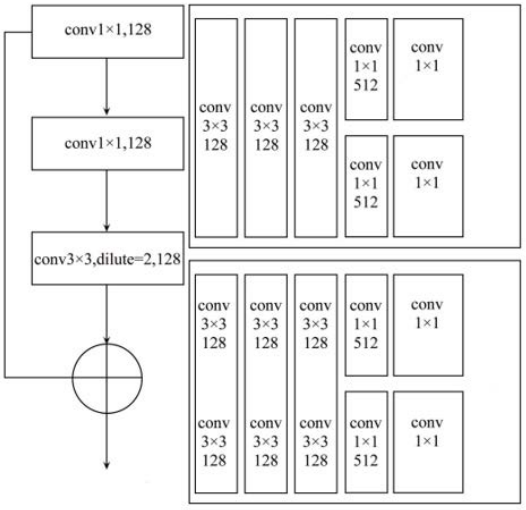


**Fig.5 Lightweight OpenPose improved structure**

The features are extracted and fed into the prediction layer of the OpenPose model to obtain the heat map of the key points of the human body and the affinity between different key points, and then fused to obtain the sequence of the human skeleton; the 7x7 convolution structure in the prediction layer is replaced by a structure consisting of the concatenation of three convolutions: a 1x7 convolution, a 7x1 convolution and a 7x7 convolution, and the outputs of the three convolutions are fused by a BN operation, and a 1x1 convolution is used before the concatenated convolution layer to compress the number of feature map channels (whole-body model) of the input concatenated convolution layer. Layer is preceded by a 1x1 convolution to compress the number of feature map channels of the input parallel convolution layer (whole body model) [2].

## 5. Pose Estimation

This chapter presents a methodology for determining three-dimensional human poses from two-dimensional pose estimation results obtained through Openpose. It evaluates both the accuracy of two-dimensional keypoint detection and the efficacy of three-dimensional keypoint triangulation under varying camera configurations.

### 5.1 Human Pose Estimation Dataset

The success of human pose estimation algorithms is attributed not only to advancements in deep learning methodologies but also to the increasing availability of sophisticated public datasets. To the best of our knowledge, there is currently no publicly available dataset specifically designed for continuous or evaluative driver monitoring using 2D human pose estimation algorithms. While manual annotation of such imagery remains a labor-intensive process, it enables direct labeling of the precise visual elements that algorithms are required to estimate. However, for 3D human pose estimation methods, this presents a considerable challenge due to the inherent loss of depth information when using conventional projective camera systems. As there are currently no publicly available datasets specifically designed for driver 3D pose estimation, Manuel Martin et al. have provided a small-scale manually annotated dataset for evaluation purposes [7].

### 5.2 Utilizing 3D Human Pose Annotation with the DRIVE&ACT Dataset

The Drive&Act dataset provides video data captured by a

ISSN 2959-6157

calibrated multi-view system comprising five near-infrared cameras and one Kinect v2. Furthermore, it offers 3D body pose estimations of the driver through triangulation of OpenPose results. The primary objective of this dataset is to facilitate fine-grained driver activity recognition.

The initial step in generating a dataset for 3D human pose estimation involves selecting appropriate frames from the Drive&Act dataset. This selection process utilizes existing 3D human pose data as a foundational reference. An analysis was conducted on each sequence within the dataset, and all instances where the positional deviation of at least one key point exceeded 10 centimeters compared to previously collected data were documented. Consequently, all frames extracted from a single sequence depict distinct postural configurations [7].

## 5.3 Three-Dimensional Pose Estimation Methodology

### 5.3.1 Multi-perspective triangulation

Triangulation primarily serves as a benchmark for subsequent depth image-based methods. Their approach employs 2D pose estimation results and performs triangulation separately for each joint. To filter outliers, the mean reprojection error (ep) for the overall body pose and the per-joint error (ej) are determined, and any joint with an absolute error exceeding 20 pixels or a relative error greater than five times ep is subsequently removed. The threshold was intentionally set at a higher value to exclusively eliminate grossly erroneous outcomes in 2D human pose estimation [7].

### 5.3.2 Deep image-based

Similar to the proposed triangulation approach, the objective is to leverage advancements in 2D human pose estimation to achieve 3D body pose estimation from depth images without requiring retraining of the 2D pose detector. One approach to achieve this objective involves detecting the driver's 2D body pose from the brightness image captured by a time-of-flight sensor, followed by determining the 3D coordinates of the 2D detection results using the corresponding depth image. A straightforward approach involves directly retrieving the depth value for each 2D body joint (u, v) from the depth image D, followed by computing the 3D joint coordinates J in the camera coordinate system using the inverse camera matrix K.

$$J_{direct} = K^{-1}\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} d(u,v) \tag{2}$$

The primary limitation of this method lies in the fact that depth images capture only the surface representation of the scene. Since actual anatomical joints are located beneath the skin, a systematic bias exists along the Z-axis. In contrast, the multi-view approach remains unaffected by this issue, as the 2D poses captured from diverse viewpoints exhibit distinct aspects of the human body. The true position of the joint can be determined by triangulating their intersection point in world space. The approach proposed by Shotton et al. [8] addresses this issue by learning offset vectors oj for each joint through training on a dataset t ∈ T.

$$J_{offset} = J_{direct} + o_j, with o_j = \frac{1}{N}\sum J_{gt} - J_{direct} \tag{3}$$

This has improved the outcomes; however, the fixed offset fails to account for variations in body shape or size. Moreover, the depth image does not provide valid z-values for occluded joints. This may result in significant measurement inaccuracies. Consequently, the offset calculated by Manuel Martin et al. is dynamically adjusted based on the input pose. This methodology enables the accurate calibration of surface joint positions to their corresponding anatomical locations within the body. Furthermore, it attempts to rectify any errors introduced by occlusion through regression-based offset correction. This is achieved via a compact and highly efficient neural network, which is applied as a post-processing step after generating the 3D pose using direct methods (see Equation (2)). This concept draws inspiration from Moon et al. [9], who employed neural networks as a post-processing mechanism to correct common errors in 2D pose estimation, such as left-right limb inversion. It is also inspired by the work of Martinez et al. [10], where a lightweight and efficient neural network can be employed in the post-processing stage to regress 3D human poses from single 2D pose estimations. Figure 7 illustrates the network architecture. The fundamental building block consists of a linear layer followed by batch normalization, a rectified linear unit (ReLU), and a dropout layer. The first module increases the feature dimensionality, while all subsequent modules maintain constant dimensions. The main network comprises two fundamental building block groups and incorporates a skip connection. The network comprises two primary components. The final layer of the network is a linear layer designed for regressing the offset corrections. It takes as input the 3D human body pose estimated through direct methods. The offset of the network output is position-independent. Therefore, the position of the input pose is normalized by subtracting the center of the input pose, which is computed as the average of all valid body joints [7].

#### 5.3.3 Two methodologies for generating displacement vectors for pose correction

The deep regression approach (depicted in the yellow segment of Figure 7) determines the output pose by applying an offset to the normalized body center point cp. This approach compels the network to reconstruct the full-body pose centered at the origin, which is subsequently translated to the correct spatial location through the addition of the body center point cp.

The deep correction method applies offsets to the corresponding input human joints (indicated by the cyan-colored components in Fig.7). This approach is analogous to the methodology described in Equation 3, but utilizes dynamic offsets that are dependent on the input pose. This approach enables more effective utilization of input poses, as they are often already highly accurate in many scenarios. Consequently, the resulting offsets are typically minimal, barring errors induced by occlusions [7].

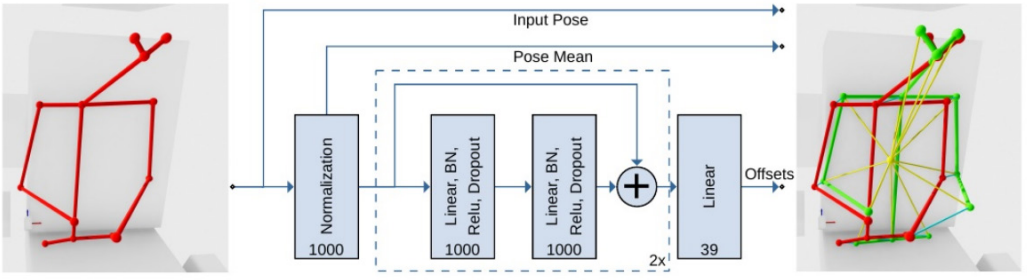


图 3. 深度修正方法的示意图。该方法以直接法得到的 3D 身体姿态作为输入（红色）。输入通过减去姿态的均值进行归一化处理。该方法通过一个小型前馈网络生成修正偏移量。我们测试了两种不同的偏移方法。一种是从用于归一化的姿态中心开始的偏移量（黄色），另一种是从输入姿态的每个关节开始的偏移量（青色）。

**Fig. 6 Schematic diagram of the depth correction method**

Fig.6 illustrates the schematic diagram of the depth correction method. This approach takes the 3D human pose obtained through direct methods as input (highlighted in red). The input is normalized by subtracting the mean pose value. A compact feedforward network is employed to generate correction offsets. Two distinct offset strategies were evaluated: one applying offsets from the pose center used for normalization (depicted in yellow), and the other applying offsets individually from each joint of the input pose (shown in cyan).

## 6. Fatigue Judgment Using OpenPose-Detected Features

### 6.1 Fatigue Features Detected by OpenPose

Table 2 lists fatigue features based on normal, mild, and severe fatigue states, and the role of key regions. Leveraging 3D images from the improved OpenPose model, this study focuses on the latter three features—nodding, head tilting, and yawning—since they are more applicable for pilots wearing sunglasses under varying lighting [4].

**Table 2. Fatigue Behavior Features**

| Feature Name | Feature Meaning |
|---|---|
| EAR Time | Longest eye closure duration |
| PERCLOS | Blink frequency |
| Nod Time | Longest head-down duration |
| Nod Count | Nod count |
| Yawn Time | Longest yawning duration |
| Yawn Count | Yawn count |
| Left Time | Longest head-tilting duration |
| Left Count | Head-tilting count |

### 6.2 Fatigue Determination from Feature Data

Method 1: Calculate the Mouth Aspect Ratio (MAR) to estimate mouth opening by computing the ratio of height to length using mouth coordinates. The MAR formula is (4), with key points shown in Figure 7.

ISSN 2959-6157

$$MAR = \frac{\| P_2 - P_6 \| + \| P_3 - P_5 \|}{2 \| P_1 - P_4 \|} \quad (4)$$

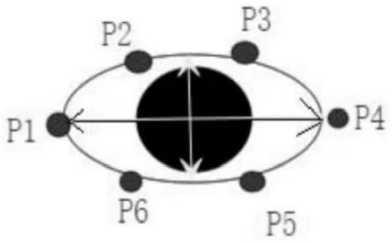


**Fig. 7 Position of Points in the MAR Formula**

A yawn is counted when MAR rises sharply (exceeding a 0.4 threshold) and then falls between consecutive frames, based on testing and analysis. Similarly, a nod is counted when the absolute difference in cosine values of the forward tilt angle between frames exceeds 0.3.

Method 2: Define features:

Definition 1: Forward tilt angle α1 is the angle between the vector formed by the neck and nose and the y-axis normal. It detects frequent nodding and prolonged head-down movements associated with fatigue. If α1 > 30° for a sustained period, fatigue is inferred (Formula 5).

$$\alpha 1 = cos^{-1} \frac{\vec{n} * \left( P_{nose} - P_{neck} \right)}{\left| \vec{n} \right| * \left| P_{nose} - P_{neck} \right|} \quad (5)$$

Definition 2: Shoulder angle α2 is the angle between vectors from the neck to the left and right shoulders. It detects large left-right head tilts during severe fatigue. If α2 > 20° with fluctuations, fatigue is inferred (Formula 6).

$$\alpha 2 = cos^{-1} \frac{\left( P_{left} - P_{neck} \right) * \left( P_{nose} - P_{neck} \right)}{\left| P_{left} - P_{neck} \right| * \left| P_{nose} - P_{neck} \right|} \quad (6)$$

Definition 3: Shoulder-neck ratio RATE is the ratio of shoulder width to neck length. Shoulder width remains stable, while neck length varies significantly with head-down movements. If RATE > 0.6 continuously, fatigue is inferred (Formula 7) [4].

$$RATE = \frac{\| Y_{nose} - Y_{shoulder} \|}{\| X_{left} - X_{right} \|} \quad (7)$$

## 7. Conclusion

In order to adapt to the evolution of commercial aviation to a single-pilot driving mode and to provide immediate feedback and safety assurance of pilot status, this paper proposes a pilot fatigue detection method based on an improved OpenPose model.

The traditional OpenPose model is difficult to be directly adapted to aviation scenarios due to the defects of data arithmetic redundancy and poor adaptability to complex lighting, etc. In order to ensure the construction of a set of intelligent detection systems that can dynamically detect, quickly analyze, and report alarms in real time, this paper proposes three major core innovations in the research through systematic improvement:

1. StyleGAN2 structure of infrared-visible fusion preprocessing: the dark state image collected by the infrared camera is converted into a highly detailed visible light image to solve the problem of facial occlusion at night or under strong light, using the sensor for dynamic aperture adjustment, combined with the cockpit light sensor to automatically adjust the camera aperture, the median filter pixels are sorted according to the RGB value, the sharpening process is enhanced for the edges, and through the two-way Generation of adversarial network, learning through the local path (eyes, nose, mouth feature points), and then synthesize the frontal face through the global path (facial contour), to solve the problem of feature offset under the driver's multi-angle sitting posture.

2. HRNet-18 lightweight network, in the same stage, the last two layers keep the two-way branching structure, and the rest of the layers are merged into a single-path structure, which is able to realize the sharing of parameters between the skeletal point confidence map PCM and the skeletal point affinity domain PAF. The original 7×7 convolution kernel was also improved by changing it to a combination of 1×1, 3×3, and 3×3 structural convolutions, and the OpenPose model was lightweighted and improved.

3. Drive&Act dataset to create a 3D skeleton to detect frequent movements during fatigue by the relative angles of the face, neck, and left and right shoulders, to provide a 3D body posture of the driver by utilizing advances in the estimation of the 2D human body posture, and to set a threshold to judge the fatigue level for fine driver activity recognition.

## Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

## References

[1] Wang M, Xiao G, Wang G. Single-Pilot Operations Technology. Journal of Aeronautics,2020, 41(04):202-220.

[2] Ruan JL,Gao P,Sun Y,et al.Research On Down Hole Personnel Behavior Detection Algorithm Based On Lightweigh OpenPose.CoalEngineering,2024,56(04):150-156.

[3] Wei YN,Zeng R,Chen X,Zhang J,Yang ZG. A Method for Detecting Cockpit Crew Fatigue.Chinese Patent:202310186133.1,2023-03-01.Application Publication

Number: CN 116503794 A, 2023-07-28.
[4] Yang J, Zhang SJ, Zhang CH, et al.A Comparative Study on Human Action Recognition Based on OpenPose.Sensors and Microsystems,2021,40(01):5-8.
[5] Ding DQ, Hu X, Meng SY, Chu DH. Fatigue Detection System, Method, Medium, and Detection Device Based on Facial Skeletal Models.Chinese Patent: 202111043280.0, 2021-09-07. Publication Authorization Number: CN 113920491 B, 2024-12-13.
[6] Phang J T S, Lim K H. Real-time multi-camera multi-person action recognition using pose estimation. Proceedings of the 3rd international conference on machine learning and soft computing. 2019: 175-180.
[7] Martin M, Voit M, Stiefelhagen R. An evaluation of different methods for 3d-driver-body-pose estimation. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021: 1578-1584.
[8] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images. CVPR 2011. Ieee, 2011: 1297-1304.
[9] Moon G, Chang J Y, Lee K M. Posefix: Model-agnostic general human pose refinement network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7773-7781.
[10] Martinez J, Hossain R, Romero J, et al. A simple yet effective baseline for 3d human pose estimation. Proceedings of the IEEE international conference on computer vision. 2017: 2640-2649.