

# ***Concurrent Recognition of Pilot Behavior via Multi-Task Learning: A Review of Architectures, Optimizations, and Challenges***

**Yiming Chen**

*School of Automation and Intelligence, Beijing Jiaotong University, Beijing, China  
wanshiruyi050401@qq.com*

**Abstract.** As civil aviation evolves toward Single Pilot Operations (SPO), the situational awareness of intelligent pilot assistance systems is critical for flight safety. Specifically, a precise, real-time understanding of pilot behavior is indispensable for compensating for the absence of "human redundancy" in the cockpit. Pilot behaviors exhibit significant spatiotemporal correlations, encompassing instantaneous operational actions (e.g., control stick manipulation) and continuous physiological poses (e.g., fatigue and observation). However, existing research often isolates these into independent single tasks, neglecting the intrinsic coupling mechanism between them in biomechanical and spatiotemporal dimensions. This results in computational redundancy, violating the stringent real-time and low-power constraints of airborne embedded platforms. To bridge this gap, this paper presents a review of concurrent recognition techniques for cockpit behaviors based on Multi-Task Learning (MTL). First, a hierarchical taxonomy of cockpit behaviors is constructed, analyzing the mechanisms underlying the correlation between actions and poses. Second, the evolution of concurrent recognition architectures is discussed in depth, with a focus on the shared feature-extraction backbone that marks the shift from 3D Convolutional Neural Networks (3D CNNs) to Spatiotemporal Transformers. Furthermore, joint optimization strategies for heterogeneous tasks, such as hard parameter sharing mechanisms and homoscedastic uncertainty weighting, are analyzed. Finally, datasets and cross-domain transfer methods are summarized. Future challenges, including environmental robustness, lightweight model deployment, and interpretability, are discussed to provide theoretical references and technical support for the construction of next-generation human-machine collaborative monitoring systems.

**Keywords:** Single Pilot Operations (SPO), Multi-Task Learning, Behavior Recognition, Spatiotemporal Transformer, Airborne Monitoring

## **1. Introduction**

The continuous rise in demand for air travel globally and the increasing pressure on airlines to lower operating costs are driving a dramatic shift in the operational mode of civil aircraft. Single Pilot Operations (SPO) has emerged as a key technological route for the construction of next-generation

large commercial aircraft, after the reduction from five-person crews to two-person crews [1]. The design objective for SPO, according to Xu et al., is to guarantee a safety level that is at least as high as that of dual-crew operations as they exist today [1]. However, successfully resolving the human factors engineering issues raised by modifications to the cockpit arrangement is necessary to meet this high safety requirement.

The pilot lacks the physical redundancy of "task sharing" and "mutual supervision" that come with conventional dual-crew arrangements when operating in SPO mode. When managing high-intensity jobs or unexpected problems, solitary pilots are vulnerable to cognitive stress in the absence of efficient auxiliary support [1]. Human factors have historically been the main cause of aviation mishaps, according to statistics on aviation safety. According to statistical study, flight crew operational errors or abnormal states (such as exhaustion or attention) are directly responsible for between 60% and 80% of aircraft mishaps [2,3]. An sophisticated cockpit monitoring system is therefore necessary to close the safety gap brought about by the "absence of a co-pilot" in SPO. Such a system must provide high-precision, real-time, all-weather sensing of pilot actions and physiological conditions using state-of-the-art computer vision to ensure aviation safety [2].

Even though prior research has made some progress in driver behavior recognition [4-7], most studies ignore the inherent spatiotemporal coupling between "continuous physiological poses" (e.g., fatigue or observation) and "instantaneous operational actions" (e.g., pushing or pulling the stick) and separate them into separate tasks. When employing numerous separate models on airborne embedded devices with constrained resources, this fragmented recognition paradigm results in significant computational duplication in addition to failing to leverage causal links between actions to boost accuracy [1,8].

In order to address the aforementioned problems, this paper reviews the current state and advancements of technologies that employ Multi-Task Learning (MTL) for the simultaneous identification of pilot actions and postures [8]. A hierarchical taxonomy of cockpit-oriented behaviors is first established, followed by an analysis of the operational logic in the cockpit, a classification of typical operational activities and physiological positions, and an explanation of the biomechanical connections between them [9]. This leads to a detailed analysis of the basic architectures for concurrent recognition, with special focus on shared feature-extraction backbones using 3D Convolutional Neural Networks (3D CNNs) and Spatiotemporal Transformers, as well as hard/soft parameter-sharing mechanisms for heterogeneous tasks [10,11]. This paper also investigates optimization and weighting strategies for joint loss functions, concentrating on the differences in training convergence speeds between posture estimation and action identification, in order to tackle the negative transfer problem in multi-task learning [12,13]. The paper ends with an overview of relevant datasets and evaluation metrics. It tackles issues such as occlusion robustness and lightweight model deployment based on an analysis of the value of cross-domain transfer from Driver Monitoring System (DMS) technologies in order to offer theoretical references and technical support for creating next-generation human-machine collaborative monitoring systems [9,14].

## 2. Behavior representation and multi-task association mechanisms

In Single Pilot Operations (SPO), an intelligent assistance system's main goal is to replace a human co-pilot's monitoring duties. The system must possess "omnidimensional perception capabilities" in order to handle complex cockpit operations. Contrary to standard single-command recognition, perception tasks in the cockpit are extremely spatiotemporal diverse and primarily rely on non-contact computer vision sensors (e.g., infrared cameras or depth cameras) to detect minute changes in the pilot in a constrained area [1].

## 2.1. Classification of pilot behaviors

The classic hierarchical model of driving behavior taxonomy divides piloting responsibilities into three categories: strategic (planning), tactical (maneuvering), and operational (control) [9]. The perceptual objectives of the intelligent cockpit can be separated into two sets of behavioral variables with distinct spatiotemporal properties when paired with categorization criteria for pilots' dangerous behaviors in civil flight [15].

The first group, referred to as immediate operational actions, includes short-duration limb motions that the pilot uses to execute flight control orders, such as pushing or pulling the side stick, pressing buttons, or utilizing the rudder pedals. In the time domain, these behaviors are distinct, highly frequent, and show a noticeable temporal reliance. Capturing the dynamic movements of limbs or face muscles in a short amount of time (often between 30 and 100 frames) is crucial for visually differentiating such activities. These include the hand-operational flow and the yawning mouth-opening action [4].

The second group consists of continuous physiological postures that characterize the long-term spatial organization of the pilot's body or mental state. Pose, as opposed to action, is spatially immobile, continuous, and low frequency. For example, the pilot's attention allocation is immediately reflected in the spatial orientation of the head (e.g., staring at the instrument panel or the HUD). On the other hand, prolonged eye closing and frequent head lowering are clear signs of exhaustion [14]. The spatiotemporal feature-extraction skills of perception algorithms are challenged in two ways by these two activities, despite their dissimilar visual presentations: they both need real-time analysis of aerial camera video streams.

## 2.2. Spatiotemporal coupling mechanism of actions and poses

Despite differences in spatiotemporal scales between operational actions and physiological poses, they do not exist in isolation biomechanically; rather, they are tightly coupled. Research indicates that a pilot's internal mental state is often a latent variable underlying external operational behavior. For example, a substantial deflection in head pose is usually accompanied by specific observation actions, whereas as a fatigue state (pose) deepens, the pilot's operational actions (e.g., pushing the stick) often become sluggish, rigid, or even deformed [5]. Therefore, the focus of learning should not rest solely on extracting single visual features, but on capturing the intrinsic correlation in which "state influences action, and action reflects state."

However, traditional single-task learning methods often sever this intrinsic link, treating action recognition and pose estimation as two independent pipelines. This approach not only ignores complementary information between tasks but also faces severe engineering challenges. Due to strict constraints on computing power, memory, and power consumption (SWaP constraints) of airborne embedded devices, directly deploying multiple independent deep neural networks would multiply the system's parameter count, resulting in severe redundancy of computational resources [16].

In contrast, the Multi-Task Learning (MTL) architecture provides an effective approach to addressing this issue. By sharing a feature extractor (Backbone) in the shallow layers of the network, MTL can significantly reduce the total model parameter count, making it more suitable for resource-constrained airborne environments. Crucially, MTL leverages inductive bias by treating the pose estimation task as an auxiliary supervision signal to regularize the model, thereby fostering more robust, general-purpose feature representations [8]. This mechanism serves as "implicit data augmentation," enabling the model to leverage pose information to constrain the feature space while

recognizing operational actions, thereby enhancing overall generalization in complex flight environments.

### 3. Evolution of recognition architectures: from 3D CNN to spatiotemporal transformer

The academic community has completed a generational leap in feature extraction backbones from Convolutional Neural Networks (CNNs) to Transformers, a paradigm shift from independent single-task networks to multi-task shared architectures, in order to achieve accurate simultaneous recognition of both actions and poses on resource-constrained airborne platforms.

#### 3.1. Hard parameter sharing

Early driver-monitoring investigations usually used a separate neural network for each task (e.g., action identification or fatigue detection). For flying embedded settings, however, this separated design is not practical since processing overhead increases exponentially with the number of operations [8]. Researchers used Multi-Task Learning's "Hard Parameter Sharing" technique to circumvent this issue. This method uses a common feature extraction backbone and many task-specific prediction heads. The prediction heads convert these characteristics to the pose regression spaces or action classification spaces, respectively, once the common backbone has acquired generic spatiotemporal representations from the input video stream [17]. In addition to drastically lowering the model's total parameters (Params) and floating-point operations (FLOPs), this technique enables the model to mine shared data across multiple tasks in the underlying feature space.

#### 3.2. Generational evolution of spatiotemporal feature modeling

The highest limit of the system's awareness of intricate cockpit motions is directly determined by the common backbone design. In recent years, popular feature extractors have changed dramatically, shifting from 3D convolution to self-attention methods.

Researchers developed the 3D Convolutional Neural Network (3D CNN) architecture in early spatiotemporal modeling to overcome the shortcomings of conventional 2D CNNs in capturing temporal information in videos. Interestingly, I3D (Inflated 3D ConvNet) learns short-term temporal aspects of motions like "pushing the stick" or "turning the head" by directly processing video clips by expanding 2D convolutional kernels along the temporal axis [18]. In order to maximize computational effectiveness even more [19,20], the SlowFast network was proposed by Feichtenhofer et al. This model has two parallel pathways: a "Fast pathway" that operates at a high frame rate to capture rapid motion (like hand trajectories) and a "Slow pathway" that operates at a low frame rate to collect spatial semantics (like facial details) [20]. The local receptive field of convolution operations fundamentally limits 3D CNNs, notwithstanding their effectiveness. This makes it challenging to capture long-range temporal dependencies needed to infer states like "deep fatigue" from minute behavioral sequences (e.g., frequent blinking) [21].

The Transformer architecture, which is based on the Self-Attention mechanism, has gradually surpassed CNNs as the de facto standard for comprehending cockpit behavior as computer vision moves into the era of huge models [10]. Transformers provide global modeling by enabling the model to explicitly compute correlations between any two spatiotemporal points in a video sequence, in contrast to 3D CNNs. The Video Swin Transformer successfully balances the trade-off between speed and accuracy among its many variations. The Video Swin Transformer presents a "Shifted Window Attention" approach to overcome the bottleneck caused by the quadratic expansion

of global attention computation with resolution, which makes typical Vision Transformers challenging to deploy in real time. Through layer-wise window shifting, it accomplishes cross-window information exchange while limiting self-attention computation to local windows [10]. In addition to lowering computational complexity to a linear level, this hierarchical architecture preserves the Transformer's capacity to capture long-term dependencies (such as fatigue evolution). Swin-T proved its effectiveness as a multi-task shared backbone by achieving a much greater recognition accuracy at a significantly lower computational cost than I3D (88 GFLOPs vs. 108 GFLOPs) in benchmarks like Kinetics-400 [10].

Furthermore, a disruptive self-supervised pre-training paradigm was presented by VideoMAE (Video Masked Autoencoders) to tackle the "data scarcity" issue in SPO [11]. This technique forces the model to rebuild the full video from the 10% of information that remains after randomly masking up to 90% of the spatiotemporal tubes in the video. Instead of only memorizing backdrop textures, the model is forced to learn incredibly robust spatiotemporal semantic properties by this exceedingly difficult pre-training assignment. According to experiments, VideoMAE offers the most promising technical route to resolving the issues of "few samples and difficult annotation" in the aviation area by achieving SOTA (State-of-the-Art) performance with only a little amount of fine-tuning data [11].

#### 4. Advanced optimization strategies: addressing negative transfer and gradient conflict

Pilot action identification (high-frequency dynamic features) and pose estimation (static geometric features) in intelligent cockpits' concurrent perception tasks show notable feature dimension variability. During combined multi-task training, this disparity frequently results in the "seesaw effect," where a sharp decline in loss for one task (such as large-amplitude stick-pushing actions) is coupled by extreme oscillation or even regression in the accuracy of another job (such as subtle tiredness features). This phenomena is essentially the "negative transfer" problem in multi-task learning, which arises from conflicts in gradient update directions and imbalances in loss magnitudes between heterogeneous tasks [8]. Dynamic loss weighting and gradient geometry correction are the two main adaptive optimization algorithms that academics have proposed to accomplish collaborative convergence of heterogeneous tasks within the shared parameter space.

Because loss values in regression tasks (like head angle prediction) are usually substantially bigger than those in classification tasks (like operational intent recognition), traditional linear weighting methods are frequently insufficient to handle uneven loss magnitudes. Kendall et al. introduced a weighting method based on homoscedastic uncertainty to eliminate the bias caused by this order-of-magnitude difference. This method introduces Bayesian probabilistic modeling by treating the loss weight for each work as a learnable noise-variance parameter ( $\sigma$ ). During training, the model automatically recognizes high-uncertainty jobs caused by intrinsic data ambiguity (e.g., image noise from sudden changes in cockpit lighting) and "penalizes" them for their contribution to the overall gradient. This approach achieves a dynamic balancing of multi-task losses and does away with the need for expensive hyperparameter grid searches [12]. Another well-liked method that dynamically determines normalization coefficients by continually monitoring the L2 norm of gradients for each job is the GradNorm methodology, which was published by Chen et al. By requiring different activities to backpropagate at equal rates, this successfully prevents simple tasks from driving network updates and guarantees that the model learns feature representations for all jobs in a balanced manner [22].

Even if the issue of loss magnitude is resolved, geometric conflicts in parameter update directions (Gradient Conflict) may still hinder model convergence. In pilot monitoring settings, the problem of

"gradient domination" is common: action identification tasks requiring large arm movements provide strong gradient signals, while fatigue detection activities requiring just minute eyelid movements produce extremely weak signals. When updated directly by superposition, strong gradients often dominate weak ones. To solve this, Yu et al. proposed an optimization method called "Gradient Surgery" (PCGrad). Before every parameter change, this algorithm calculates the angle between task gradients. When an obtuse angle (greater than 90 degrees) is found, indicating competing optimization directions, the gradient of one work is projected onto the normal plane of the other job. Essentially, by removing the conflicting gradient components, this orthogonalization procedure ensures that the shared parameters update downward for each job while preserving meaningful information [13].

Furthermore, in response to the long-tail distribution of training difficulty brought on by environmental noise in SPO, such as camera shake from turbulence or low light levels, Liang et al. proposed the BMTL (Balanced Multi-Task Learning) framework, which introduces a dynamic difficulty-perception mechanism. This acts as an attention mechanism in later training stages, allowing the model to automatically detect "hard samples" based on the historical loss trends and focus on difficult clips with poor lighting or blurred actions, significantly improving perception robustness under adverse flight conditions [23].

Among these, homoscedastic uncertainty weighting is considered the optimal choice for contemporary avionics scenarios. Unlike GradNorm or PCGrad, which require complex geometric projections or frequent gradient norm calculations, uncertainty weighting adds few learnable parameters and imposes almost no additional computational resource consumption (SWaP) during inference, meeting the stringent reliability requirements of aviation airworthiness standards [12].

## 5. Datasets, evaluation, and cross-domain transfer

High-quality annotated data is the cornerstone of improving deep learning model performance. However, in the aviation domain, due to strict privacy protection regulations (such as CVR-related laws) and flight safety standards [24], there is currently no publicly available large-scale dataset of pilot behavior in academia [2,9]. Facing this "data silo" dilemma, existing solutions mainly follow a dual-track strategy of "simulator generation as the primary source and cross-domain transfer as the auxiliary." On the one hand, using high-fidelity flight simulators to build derivative datasets is the preferred approach for obtaining in-domain prior knowledge. For instance, Liu et al. used a civil aviation simulation visual system to collect manipulation sequences under complex lighting to construct a keypoint dataset [25], while Man et al. captured typical abnormal behavior video streams by presetting violation scripts [3]. Such data can effectively cover "long-tail samples" that are extremely difficult to capture in real flights. On the other hand, to compensate for the lack of diversity and scale in simulation data, leveraging rich resources from the automotive Driver Monitoring System (DMS) field for transfer learning has become a key path for building intelligent cockpit perception systems [9]. This is based on the high biomechanical commonality between the two: operators in both cars and aircraft cockpits are in a restricted seated space, and the physiological external manifestations of fatigue (e.g., yawning, closed eyes) and distraction (e.g., gaze deviation) possess cross-scenario consistency [14]. Therefore, mainstream DMS datasets such as StateFarm (focusing on distraction) and Drive&Act (focusing on fine-grained interaction actions) are often used as source domain-domain data for training pilot behavior recognition models [17].

To successfully adapt models trained for automotive drivers to pilot scenarios, researchers typically employ a "pre-training & fine-tuning" transfer strategy. Specifically, the shared backbone network is first pre-trained on massive DMS video data (e.g., over 9 million frames of multi-view

video in Drive&Act [9,17]) to learn general human motion features (e.g., hand trajectories, facial contours) and spatiotemporal semantic representations. Subsequently, the shallow general feature extraction layers of the network are frozen, and only the deep task-specific heads are fine-tuned using a small amount of pilot simulation data. This approach avoids overfitting brought on by direct training on small-sample aircraft data by efficiently using the richness of source-domain data to establish model weights. Furthermore, addressing the "domain shift" challenge unique to aviation scenarios—such as strong lighting contrast during high-altitude cruising [2] and image blurring caused by turbulence—self-supervised architectures like VideoMAE have demonstrated unique advantages. By using masked reconstruction tasks to force the model to focus on the high-level semantics of objects rather than low-level pixel distributions, VideoMAE significantly enhances the model's robustness in cross-domain scenarios, maintaining high recognition accuracy even in unseen cockpit lighting environments [11].

Table 1. Performance comparison of mainstream spatio-temporal behaviour recognition algorithms on the Kinetics-400 benchmark

Method	Backbone	Top-1	GFLOPs	Param
I3D [10,20]	ResNet-50	72.1%	108 G	25.0 M
SlowFast [20]	ResNet-50	75.6%	36.1 G	34.4 M
Video Swin-T [10]	Swin-Tiny	78.8%	88.0 G	28.2 M
VideoMAE [11]	ViT-Base	81.5%	180 G	87 M
M2DAR [17]	Hierarchical-ViT	N/A	94 G	42 M

## 6. Challenges and future trends

The deployment of cockpit monitoring technologies based on multi-task deep learning from laboratories to real Single Pilot Operations (SPO) routes still faces many obstacles, from algorithm robustness to legal airworthiness certification, despite the promising performance shown in experimental settings.

First, regarding technical deployment, the highly unstructured nature of the real flight environment constitutes a severe perception barrier. The reliability of vision-based algorithms may be jeopardized by the sharp dynamic range difference between intense direct sunlight during high-altitude cruising and extremely low illumination during night flights, as well as self-occlusion phenomena brought on by pilots wearing oxygen masks, sunglasses, or hand operations. The main difficulty for all-weather monitoring is improving the models' resilience to algorithm interference under high lighting conditions and partial occlusion because the majority of existing models are trained on relatively ideal simulated data. At the same time, airborne avionics must meet strict airworthiness standards for weight, power, and heat dissipation (SWaP restrictions) [1]. Despite achieving great identification accuracy, current high-performance Transformer models are not appropriate for straight onboard deployment due to their enormous parameter counts. Thus, model lightweighting will unavoidably be a trend in future study. For engineering implementation, it is essential to use methods like Knowledge Distillation and Network Pruning to condense models to a size that edge computing devices can handle while preserving high accuracy [16,26,27].

Second, the aviation industry's intense quest for system interpretability, which is closely related to system airworthiness certification and the development of human-machine trust, is inherently at odds with the "black box" character of deep learning models [1]. The lack of obvious logical

reasoning pathways (i.e., "why the system believes the pilot is fatigued") in current end-to-end networks, despite their ability to reliably output fatigue levels, is unacceptable to regulatory authorities when it comes to choices involving flight safety. Consequently, one of the major trends in this subject will be the development of Explainable Artificial Intelligence (XAI). To intuitively show the neural network's decision basis to pilots or ground monitors, researchers must implement methods like Attention Visualization or Class Activation Mapping (Grad-CAM) [9]. For instance, they can highlight the hand operation trajectories or eyelid closure regions that the model focuses on. In order to achieve the strict requirements for software Design Assurance Level (DAL) in airworthiness certification, algorithms must have "white-box" transparency in order to genuinely foster human pilots' faith in intelligent assistance systems [1].

Lastly, the adoption of the SPO mode involves a thorough rebuilding encompassing aviation legislation and ethical obligations in addition to a technological revolution. With explicit legal definitions for crew configuration, duty period constraints, and task division, current civil aviation operating standards (e.g., CCAR-121) are built around dual-crew operations [24]. The legal boundaries between human and machine rights and obligations must be defined when intelligent monitoring systems partially replace co-pilot duties. For instance, is the pilot's supervisory error or the algorithm's bad decision-making to blame for a safety incident? This uncertainty in responsibility attribution continues to be the biggest non-technical barrier to the commercial deployment of SPO. Therefore, future research must go beyond a single technological perspective and promote the cross-integration of artificial intelligence, ethics, and jurisprudence in order to jointly develop next-generation airworthiness standards and operational specifications that are tailored to human-machine collaboration [1].

## 7. Conclusion

The evolution of the Single Pilot Operations (SPO) mode marks the entry of civil aviation into a new era of human-machine collaboration, and building an intelligent cockpit monitoring system with high situational awareness capabilities is the key technological means to fill the void of "artificial redundancy" caused by the absence of a co-pilot. By examining the spatiotemporal characteristics of pilot behaviors, this paper demonstrates the intrinsic coupling between action recognition and pose estimation. It points out that adopting a multi-task learning architecture with a shared backbone—specifically, a hard-parameter-sharing network based on the Video Swin Transformer—is the optimal solution for balancing airborne computing bottlenecks and feature heterogeneity. Addressing the challenges of aviation data scarcity and cross-domain transfer, the dual-track strategy integrating high-fidelity simulation and DMS transfer learning has been proven to be an effective path for enhancing model generalization capabilities, while the introduction of homoscedastic uncertainty weighting effectively resolves gradient conflicts between heterogeneous tasks. Although the industry still faces multidimensional challenges such as insufficient environmental robustness, difficulties in lightweight model deployment, and complex airworthiness certification, with the maturity of self-supervised learning paradigms, breakthroughs in Explainable AI (XAI) technologies, and the perfection of relevant regulatory systems, intelligent perception technology is bound to gradually break through the "black box" trust barrier. It will become an indispensable "digital co-pilot" in future human-machine collaborative systems, providing solid theoretical and technical support for the safe operation of next-generation civil aviation.

## References

- [1] Xu Wei, Chen Yong, Dong Wenjun, et al. "Status and prospect of human factors engineering research on single pilot operations for large commercial aircraft," *Advances in Aeronautical Science and Engineering*, vol. 13, no. 1, pp. 1-18, 2022. DOI: 10.16615/j.cnki.1674-8190.2022.01.01.
- [2] Man Yongzheng. "Research on key technologies of pilot typical abnormal behavior recognition based on deep learning," Ph.D. dissertation, Civil Aviation Flight University of China, 2023. DOI: 10.27722/d.cnki.gzgmh.2023.000298.
- [3] Chen Nongtian, Man Yongzheng, Ning Weifeng, et al. "An abnormal driving behavior monitoring method of pilot based on deep learning," *Journal of Safety and Environment*, vol. 22, no. 1, pp. 249-255, 2022. DOI: 10.13637/j.issn.1009-6094.2021.1217.
- [4] Lu Y, Liu C, Chang F, et al. JHPFA-Net: Joint head pose and facial action network for driver yawning detection across arbitrary poses in videos [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2023, 24(11): 11850-11863.
- [5] Hu Yaocong. "Research on driver behavior and fatigue recognition method based on deep learning," Master's thesis, Southeast University, 2021. DOI: 10.27014/d.cnki.gdnau.2021.003939.
- [6] Bai J, Yu W, Xiao Z, et al. Two-stream spatial-temporal graph convolutional networks for driver drowsiness detection [J]. *IEEE Transactions on Cybernetics*, 2021, 52(12): 13821-13833.
- [7] Deng W, Wu R. Real-time driver-drowsiness detection system using facial features [J]. *Ieee Access*, 2019, 7: 118727-118738.
- [8] Alzahrani M, Wang Q, Liao W, et al. Survey on multi-task learning in smart transportation [J]. *Ieee Access*, 2024, 12: 17023-17044.
- [9] Bouhsissin S, Sael N, Benabbou F. Driver behavior classification: a systematic literature review [J]. *IEEE Access*, 2023, 11: 14128-14153.
- [10] Liu Z, Ning J, Cao Y, et al. Video swin transformer [C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 3202-3211.
- [11] Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training [J]. *Advances in neural information processing systems*, 2022, 35: 10078-10093.
- [12] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 7482-7491.
- [13] Yu T, Kumar S, Gupta A, et al. Gradient surgery for multi-task learning [J]. *Advances in neural information processing systems*, 2020, 33: 5824-5836.
- [14] Kamboj M, Kadian K, Dwivedi V, et al. Advanced detection techniques for driver drowsiness: a comprehensive review of machine learning, deep learning, and physiological approaches [J]. *Multimedia Tools and Applications*, 2024, 83(42): 90619-90682.
- [15] Wang Lei, Wei Zixin, Zou Ying. "Classification management method of errors and violations in civil aviation pilot unsafe behaviors," *China Safety Science Journal*, vol. 34, no. 12, pp. 8-15, 2024. DOI: 10.16265/j.cnki.issn1003-3033.2024.12.0281.
- [16] Tang Chulin. "Research on driver behavior recognition method based on multi-modal information fusion," Master's thesis, Hunan University, 2023. DOI: 10.27135/d.cnki.ghudu.2023.003508.
- [17] Ma Y, Yuan L, Abdelraouf A, et al. M2DAR: Multi-view multi-scale driver action recognition with vision transformer [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023: 5287-5294.
- [18] Zhao L, Wang Z, Zhang G, et al. Driver drowsiness recognition via transferred deep 3D convolutional network and state probability vector [J]. *Multimedia Tools and Applications*, 2020, 79(35): 26683-26701.
- [19] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]//*European conference on computer vision*. Cham: Springer International Publishing, 2016: 20-36.
- [20] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition [C]//*Proceedings of the IEEE/CVF international conference on computer vision*. 2019: 6202-6211.
- [21] Wang X, Girshick R, Gupta A, et al. Non-local neural networks [C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 7794-7803.
- [22] Chen Z, Badrinarayanan V, Lee C Y, et al. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks [C]//*International conference on machine learning*. PMLR, 2018: 794-803.
- [23] Liang S, Deng C, Zhang Y. A simple approach to balance task loss in multi-task learning [C]//*2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021: 812-823.

- [24] Civil Aviation Administration of China (CAAC). Large Aeroplane Air Carrier Certification: CCAR-121-R7 [S]. Beijing: CAAC, 2021.
- [25] Liu Hao, Sun Youchao, Wu Honglan, et al. "Keypoint detection method of civil aircraft pilot in complex lighting environment, " Journal of Beijing University of Aeronautics and Astronautics, vol. 51, no. 10, pp. 3471-3481, 2025. DOI: 10.13700/j.bh.1001-5965.2023.0566.
- [26] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.
- [27] Mehta S, Rastegari M. Light-weight, general-purpose, and mobile-friendly vision transformer [J]. arXiv e-prints, 2021: arXiv: 2110.02178, 2021.