electronics

MDPI

Article

A Dual-Branch ST-GCN System for Joint Recognition of OOW Unsafe Behaviors and Facial Fatigue Features

Rui Qi, Shengwei Xing *, Kairen Chen, Zijian Zhang and Xiaoyu He

College of Navigation, Dalian Maritime University, Dalian 116026, China; qr_dlmu@dlmu.edu.cn (R.Q.); 2201902802ckr@dlmu.edu.cn (K.C.); 24zzj@dlmu.edu.cn (Z.Z.); hxy_2003@dlmu.edu.cn (X.H.)
* Correspondence: xingsw@dlmu.edu.cn

Abstract

The Officer on Watch (OOW) is critical to ensuring the safety of the vessel, cargo, and crew during navigation. To reduce maritime accidents caused by unsafe behaviors or fatigue, this paper proposes a dual-branch detection system based on Spatial–Temporal Graph Convolutional Networks (ST-GCN): BODY-ST-GCN for pose-based behavior recognition and FACE-ST-GCN for facial state analysis. For spatial modeling, a Triple Graph Fusion (TGF) strategy is introduced to integrate static, adaptive, and attention graphs, enhancing the representation of skeletal and facial keypoints. For temporal modeling, BODY-ST-GCN incorporates a Three-Scale Parallel Temporal Convolutional Network (TSP-TCN) to capture multi-scale motion dynamics, while FACE-ST-GCN uses a Temporal Adaptive Module (TAM) to extract stable facial state features. Furthermore, a joint risk classification mechanism categorizes OOW duty states into four hierarchical levels: Safe, Early Fatigue Warning, High Fatigue Risk, and Emergency. This mechanism enables continuous, real-time monitoring and dynamic assessment. Experiments demonstrate that BODY-ST-GCN and FACE-ST-GCN achieve macro average precisions of 0.969 and 0.947, respectively, outperforming the baseline ST-GCN by 6.4% and 14.9%, providing reliable technical support for onboard safety management.

Keywords: OOW; OpenPose; ST-GCN; behavior and fatigue detection; maritime safety

1. Introduction

Shipping is the main way of global cargo transportation. However, the occurrence of maritime accidents not only threatens the global economy and ecology, but also pose a severe risk to human safety. In recent years, maritime safety accidents have occurred frequently. According to the Maritime Safety Report published by EMSA in 2023, from 2014 to 2022, a total of 5941 maritime accidents occurred, resulting in 6781 injuries. Meanwhile, 59.1% of these accidents involved human behavior, and 50.1% of all contributing factors were attributed to this cause. When combining human behavior events and contributing factors, 80.7% of maritime accidents were linked to human factors [1]. Studies have confirmed that human factors are a major cause of maritime accidents [2]. The Officer on Watch (OOW) is responsible for ensuring the safety of the vessel's navigation and monitoring the vessel's status. Unsafe behavior exhibited by OOW is one of the key causes of maritime accidents. Although the International Maritime Organization (IMO) MSC.128(75) resolution mandates the installation of the Bridge Navigational Watch Alarm System (BNWAS) on vessels [3], BNWAS relies only on reset operations to confirm the presence of OOW and cannot identify OOW's behavior or the specific cause of the alarm. With the continuous development of

Check for updates

Received: 1 March 2026
Revised: 20 April 2026
Accepted: 23 April 2026
Published: 27 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

deep learning technology and its widespread application in the maritime field [4–6], using intelligent technologies to monitor OOW's behavior and status in real-time is crucial for reducing maritime accidents caused by human factors.

Human pose estimation algorithms provide the foundation for downstream tasks such as pose tracking and behavior prediction [7,8]. These algorithms can detect and track keypoints on the body, such as the head, shoulders, and torso, in real time, analyzing their spatial position to accurately capture human movement. Using human pose estimation algorithms allows for real-time monitoring of OOW's specific behavior and provides early warnings of risks, effectively improving work efficiency. However, human pose estimation algorithms can only provide keypoint information of the body and lack the ability to recognize specific behaviors. Therefore, selecting an appropriate behavior recognition algorithm is crucial for detecting the unsafe behavior of OOW.

Currently, human action recognition based on Spatial–Temporal Graph Convolutional Networks (ST-GCN) [9] has become a research hotspot. ST-GCN effectively captures the dynamic changes in keypoints in both spatial structure and time series by integrating the spatial–temporal information of keypoints during action changes, enabling the detection of human actions, facial actions, and more. However, the original ST-GCN has certain limitations. First, its spatial structure is fixed, which makes it less flexible when dealing with complex human motion variations in real-world scenarios. Second, the original ST-GCN can only model the temporal dimension at a fixed scale, limiting its ability to capture multi-scale temporal information and restricting its temporal modeling capability. Moreover, it lacks an effective facial topological structure as the foundation graph for the network, which limits its ability to model and recognize facial features.

Meanwhile, existing studies on driver behavior detection and fatigue monitoring are mostly focused on single-task scenarios, lacking a unified framework that can jointly model both behavioral patterns and facial fatigue states. Such a separated modeling strategy makes it difficult to comprehensively characterize the driver's overall status in real-world watchkeeping environments, thereby limiting the practical applicability of these systems. Therefore, there is a clear need to develop a unified and adaptive framework capable of simultaneously modeling human behaviors and facial fatigue features while enhancing feature representation in both spatial and temporal dimensions.

To address these issues, this paper proposes a dual-branch ST-GCN framework for recognizing unsafe driver behaviors and facial fatigue features. The proposed method first employs the OpenPose algorithm to extract both body and facial keypoint information. Subsequently, two separate ST-GCN networks are constructed for unsafe behavior recognition and facial fatigue feature detection, respectively. In addition, a dedicated facial topology is designed for facial feature modeling, and targeted improvements are introduced to both ST-GCN networks, enabling effective detection of unsafe behaviors and facial fatigue states.

The main contributions of this paper can be summarized as follows:

- This paper addresses the task of monitoring unsafe behaviors of OOW by constructing a BODY-ST-GCN detection branch. Considering the complex spatial structural relationships and the diverse temporal dynamics of human actions, a Triple Graph Fusion (TGF) strategy is introduced in spatial modeling to enhance the representation capability of keypoint relationships. In temporal modeling, a Three-Scale Parallel Temporal Convolutional Network (TSP-TCN) is designed to effectively capture multi-scale temporal features of different behaviors, thereby improving the recognition ability for complex action patterns.
- To address the characteristics of facial fatigue in OOW, which exhibit relatively small variations and unstable durations, this paper designs a FACE-ST-GCN detection branch. This branch constructs a facial topological structure based on facial keypoints

extracted by OpenPose as the foundation of the graph model and incorporates a Temporal Attention Module (TAM) in temporal modeling to enhance the model's ability to identify key frames, thereby better capturing fatigue-related facial changes.
- For real-world ship bridge scenarios, the proposed method for recognizing unsafe behaviors and fatigue states of OOW has been validated through practical applications. Experiments are conducted on both public datasets and real-world collected datasets. The results demonstrate that the proposed method can effectively improve the recognition performance of unsafe behaviors and fatigue states, indicating strong potential for practical application.

The remainder of the paper is organized as follows: Section 2 reviews the related research on unsafe behavior detection in crew members, human posture estimation, and action recognition; Section 3 provides a detailed description of the proposed model and the mechanism for detecting OOW unsafe and fatigue behaviors; Section 4 presents the experiments and results analysis, validating the effectiveness of the proposed improvements and the overall performance of the model through ablation and comparative experiments, with volunteer participants acting as OOW in a real ship's bridge for unsafe behavior detection; Section 5 concludes the paper, summarizing the research findings, analyzing the model's strengths and limitations, and outlining directions for future optimization.

2. Related Work

2.1. Research on Unsafe Behaviors of Crew

The IMO adopted the casualty investigation code in 1999, which proposed a systematic approach to investigating human factors in maritime accidents and defined unsafe behavior as "A departure from acceptable or desirable practice on the part of an individual or group of individuals that can result in unacceptable or undesirable results" [10]. In recent years, fatigue monitoring technology based on electroencephalography (EEG) has been widely applied in the maritime field [11–13]. This technology typically uses sensors to collect EEG signals and other physiological data from crew members. After feature extraction and filtering, the crew member's status is identified through machine learning classification models or other methods. However, when monitoring the crew's status based on physiological data such as EEG, practical applications face challenges due to factors such as individual differences and device limitations.

To further monitor the specific behaviors and activity states of OOW, Youn et al. [14] proposed a crew lookout behavior classification method based on Kinect sensors, while Chen et al. [15] introduced a Wi-Fi-based method for monitoring OOW activity states and quantities. However, both methods face issues such as limited hardware deployment and instability in complex environmental adaptability. With the advancement of deep learning technology, research applying it to detect unsafe behaviors of crew members has gradually increased. For example, Lin et al. [16] proposed the YOLO-SD method to detect whether crew members performing high-altitude operations wear safety ropes, Zhao et al. [17] used YOLOv4 to identify unsafe mooring and unmooring behaviors of crew members, and Liu et al. [18] combined BlazePose and LSTM to detect crew member falls. However, most of these studies focus on detecting crew behaviors in other operational areas of the ship, and there is limited research on using deep learning techniques for detecting OOW behaviors inside the bridge.

Therefore, this paper combines human posture estimation algorithms and ST-GCN to implement behavior detection of OOW during the duty process.

2.2. Human Pose Estimation Algorithm

Human pose estimation algorithms can be categorized into single-person pose estimation and multi-person pose estimation. Compared to multi-person pose estimation, single-person pose estimation has lower algorithm complexity and focuses on individual motion analysis. However, its accuracy in keypoint detection is lower when occlusions are present. Multi-person pose estimation is currently divided into two main types: Top-down and Bottom-up approaches. The Top-down approach involves two stages: first, a human detection algorithm detects humans in the image, and then keypoints are annotated on the detected human bounding boxes. For example, the Cascaded Pyramid Network (CPN) by Chen et al. [19] and the Simple Baselines algorithm by Xiao et al. [20] both adopt a Top-down detection approach.

However, the Top-down approach heavily relies on the accuracy of human detection bounding boxes, which fail to effectively capture human poses when individuals are occluded. The Bottom-up approach is an end-to-end detection approach, which first detects and annotates keypoints on humans, and then uses the relationships between the keypoints to identify the corresponding individuals. For instance, PifPaf proposed by Kreiss et al. [21] detects the positions of human keypoints by designing a Part Intensity Field (PIF) and then uses integer linear programming to form the final human pose from the candidate parts; HigherHRNet is a model proposed by Cheng et al. [22] based on High-Resolution Net (HRNET), which achieves higher resolution detection performance that solves the issue of image scale variations during detection through heatmap deconvolution, thereby enhancing the recognition of complex human poses and subtle body movements; OpenPose is a real-time multi-person 2D pose estimation algorithm based on part affinity fields, proposed by Cao et al. [23]. Compared to PifPaf and HigherHRNet, OpenPose performs better in terms of keypoints connection efficiency, real-time performance, and multi-person detection capabilities. Moreover, OpenPose covers a broader range of keypoints detection and is capable of detecting the human skeleton, hands, face, and feet with a total of 135 keypoints, thereby providing a more comprehensive human pose estimation task. Therefore, this paper uses OpenPose as the base model to achieve comprehensive detection of OOW body and facial keypoints.

2.3. Human Action Recognition

Human action recognition (HAR) aims to understand human action and classify actions based on motion trajectories. Neural networks can extract features at deeper levels, and action recognition using skeleton information that leverages variations in human keypoints has become a widely adopted approach. This method utilizes skeleton information obtained through human pose estimation algorithms or motion capture devices and feeds it into recognition algorithms for action classification. Since it focuses on the intrinsic motion features of the human body, it is not affected by background information in images and typically exhibits stronger robustness and generalization ability [24]. At present, action recognition methods using neural networks can be categorized into four main types: sequence modeling methods based on RNNs, spatiotemporal feature behavior methods based on CNNs, graph structure modeling methods based on GCNs, and global modeling methods based on Transformers.

RNNs have advantages in handling sequential data, and skeleton data are typically converted into time series and then fed into RNNs or LSTMs for action recognition [25–27]. However, due to their inherent architectural limitations, RNNs exhibit insufficient spatial modeling capability. CNNs achieve action recognition through the transformation of skeleton data into pseudo-images, followed by feature action [28–30]. Transformers [31] rely entirely on the attention mechanism to model the relationships among keypoints. For

example, Qiu et al. [32] proposed the Spatio-Temporal Tuples Transformer (STTFormer), which employs a spatiotemporal tuple self-attention module to jointly model human keypoints and explicitly capture their coordinated features. Wu et al. further developed the FreqMixFormerV2 [33] and UniSTFormer [34] models. The former reduces the number of attention modules and introduces frequency operators to improve computational efficiency while enhancing the discrimination of subtle motions. The latter adopts a unified spatiotemporal modeling strategy and integrates both global and local features, enabling a more comprehensive representation of spatiotemporal information. However, existing Transformer-based skeleton action recognition methods rely entirely on attention mechanisms to model relationships among keypoints and lack the incorporation of inherent skeletal structural priors, making them susceptible to noise in keypoints detection.

Due to the unique structural characteristics of skeleton data, simply mapping them into coordinate vector sequences or pseudo images often results in the loss of their inherent structural information. Consequently, GCNs designed for learning and reasoning on graph data have become a popular research direction in action recognition from skeleton data. Yan et al. [9] first proposed a unified spatiotemporal modeling approach for skeleton sequences, introducing the ST-GCN, which can automatically capture both the spatial configurations and dynamic variations in skeleton sequences. Subsequent research on skeleton-based behavior recognition with GCNs has primarily focused on improving ST-GCN, such as the Two-Stream Adaptive Graph Convolutional Networks (2S-AGCN) proposed by Shi et al. [35] and the Actional–Structural Graph Convolutional Networks (AS-GCN) proposed by Li et al. [36]. In this paper, we optimize ST-GCN by proposing a two-branch ST-GCN framework designed for recognizing OOW unsafe behaviors and facial fatigue features associated with fatigue.

3. Methods

3.1. Algorithm Structure

As illustrated in Figure 1, the proposed method for joint recognition of OOW unsafe behaviors and facial fatigue states consists of three stages: human keypoints extraction, dual-branch feature modeling, and risk state assessment.

In the input stage, the collected OOW watchkeeping videos are processed using the OpenPose algorithm to simultaneously extract human skeletal keypoint sequences and facial keypoint sequences. The skeletal keypoints are used to represent body movements, while the facial keypoints capture variations in facial states. In the feature modeling stage, a dual-branch ST-GCN framework is constructed to model body behaviors and facial states, respectively. For the body behavior detection branch (BODY-ST-GCN), a TGF strategy is incorporated into the GCN to enhance the relational representation among keypoints. In the temporal domain, a TSP-TCN module is designed to replace the original TCN, thereby improving the model's ability to capture behavioral variations across different temporal scales. For the facial state detection branch (FACE-ST-GCN), the TGF strategy is similarly introduced into the GCN to strengthen spatial modeling capability. In addition, a TAM is embedded into the TCN to improve the modeling of subtle facial changes at critical moments. Finally, in the decision fusion stage, based on predefined risk assessment rules, the behavior recognition results from BODY-ST-GCN and the facial state recognition results from FACE-ST-GCN are jointly analyzed and mapped to OOW risk levels, yielding the final assessment outcome.

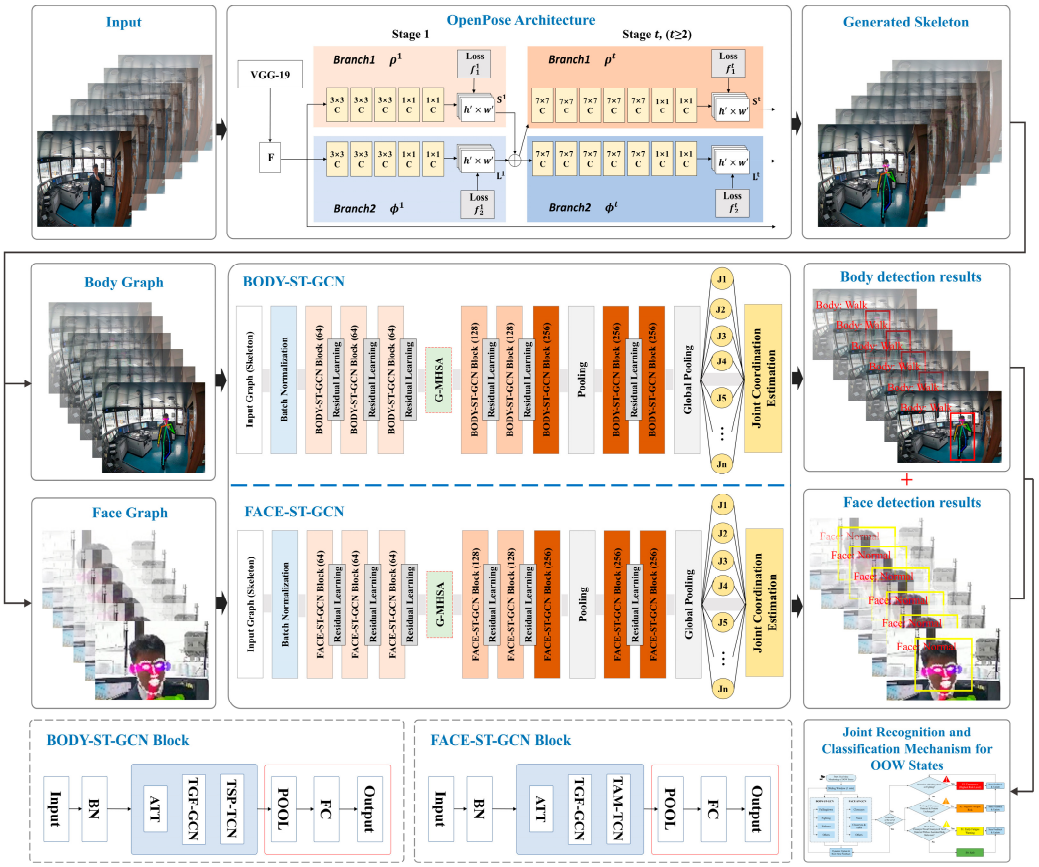


Figure 1. The structure of the OOW unsafe behaviors and the facial fatigue features detection algorithm.

3.2. OOW Unsafe Behaviors Selection and Fatigue State Judgment Criteria

In this paper, three types of body behaviors that are potentially associated with risk during OOW watchkeeping are selected for analysis, namely "sitdown," "fallingdown," and "fighting." The selection of these behaviors is primarily based on their relevance to safety risks in real navigation scenarios. During watchkeeping, prolonged "sitdown" posture may lead to reduced alertness of the OOW, which can indirectly affect their operational performance and decision-making ability in navigation. In contrast, "fallingdown" and "fighting" are considered more direct high-risk behaviors. Under rough sea conditions, vessel motion induced by wind and waves may cause significant instability, increasing the likelihood of the OOW losing balance and falling, which poses a threat to personal safety and may result in a loss of control over the vessel. In addition, after extended periods of watchkeeping, the combined effects of stress and fatigue may lead to emotional instability, potentially triggering conflict behaviors. Although such situations occur with relatively low probability, once they arise, they may pose serious risks to both human safety and vessel navigation.

Beyond body posture changes, the fatigue state of OOW is also a critical factor that warrants attention. Facial variations are widely recognized as key indicators for assessing human fatigue. Under fatigued conditions, the most notable changes occur in two facial regions: the eyes and the mouth. Among these, the eye region provides a direct and reliable reflection of severe fatigue. The percentage of eyelid closure over the pupil over time (PERCLOS) has been widely acknowledged as one of the most effective metrics for fatigue assessment. PERCLOS is defined as the proportion of time within a given interval during which the eyelids cover the pupil, and its evaluation typically relies on several well-established threshold criteria:

- EM: The threshold is set when the proportion of the pupil covered by the eyelid exceeds 50%, used as a reference standard.
- P70: The threshold is set when the proportion of the pupil covered by the eyelid exceeds 70%, used as a reference standard.
- P80: The threshold is set when the proportion of the pupil covered by the eyelid exceeds 80%, used as a reference standard.

In shipboard operations, certain behaviors of OOW, such as looking downward to read documents or operate instruments, may also lead to partial or complete eyelid closure, thereby introducing potential ambiguity in fatigue assessment. To alleviate this issue, a temporal constraint is incorporated into the fatigue judgment process, with emphasis placed on identifying sustained eye closure rather than instantaneous eye states. Specifically, fatigue is assessed using the P70 criterion of the PERCLOS metric, where a one-minute evaluation window is adopted. If the proportion of eyelid closure exceeds 70% within this interval, the behavior is defined as prolonged eye closure, indicating that OOW is in a fatigued (sleep) state.

In addition, this paper considers the "yawning" behavior of the personnel. Yawning is generally regarded as being associated with fatigue and can, to some extent, reflect changes in an individual's state. However, given the considerable inter-individual variability in yawning behavior, although it is recognized in the subsequent facial state detection, it is not used as a criterion for determining whether a person is in a fatigued state. Instead, it is treated as an auxiliary feature to support the analysis of fatigue-related conditions of the OOW.

3.3. Joint Recognition and Classification Mechanism for OOW Risk States

In this paper, the duty states of the OOW are classified into four levels: S0 (Safe), S1 (Early Fatigue Warning), S2 (High Fatigue Risk), and S3 (Emergency). The Safe state corresponds to normal duty conditions without evident fatigue or unsafe behaviors. The Early Fatigue Warning state indicates the emergence of initial fatigue-related signs. The High Fatigue Risk state is characterized by the presence of fatigue features within a specified time window, accompanied by unchanged posture. The Emergency state denotes the occurrence of critical hazardous behaviors. The definitions and corresponding behavioral criteria of these four risk states are presented in Table 1.

Table 1. Four risk states and their corresponding behavioral criteria.

State	Description	Body Condition	Face Condition
S0	Safe	No fallingdown or fighting detected, and no sustained body behavior	Normal facial state, with no closeeyes, yawn, or closeeyes_yawn detected
S1	Early Fatigue Warning	No sustained body behavior and no fallingdown or fighting detected	Closeeyes, yawn, or closeeyes_yawn detected, but the criterion for prolonged close-eyes is not met
S2	High Fatigue Risk	Posture remains unchanged within the specified time window	Prolonged closeeyes detected
S3	Emergency	Fallingdown or fighting detected	Any facial state

Based on the above state definitions, the decision rules for OOW risk state classification are further established as follows:

- S0 (Safe): When no unsafe or fatigue-related behaviors, such as fallingdown, fighting, or prolonged close-eyes, are detected, the OOW is classified as being in the S0 (Safe) state.
- S1 (Early Fatigue Warning): When fatigue-related features such as closeeyes or yawn are detected, but the criterion for prolonged closeeyes is not satisfied, and no sustained body movement is observed, the OOW is classified as being in the S1 (Early Fatigue Warning) state.
- S2 (High Fatigue Risk): When prolonged closeeyes is detected within a unified time window, and the OOW maintains an unchanged posture during this time window, the OOW is classified as being in the S2 (High Fatigue Risk) state.
- S3 (Emergency): When hazardous behaviors such as fallingdown or fighting are detected, the OOW is directly classified as being in the S3 (Emergency) state, regardless of the facial condition.

To ensure consistency in the identification of high fatigue risk states, the time window used for S2 state determination is set to be the same as that adopted for prolonged close-eyes detection, both defined as 1 min. Among the four states, S3 (Emergency) is assigned the highest priority to prevent interference from other fatigue-related features in emergency risk recognition. Considering the temporal continuity of OOW duty states, a sliding time window is employed to dynamically update the risk state classification. The window length is set to be consistent with that used for S2 state determination, i.e., 1 min, and is updated in real time with a fixed step size along the time sequence. This enables continuous monitoring and dynamic assessment of the OOW risk state. The overall workflow of the dual-branch ST-GCN framework for multimodal behavior detection and hierarchical risk assessment is shown in Figure 2.

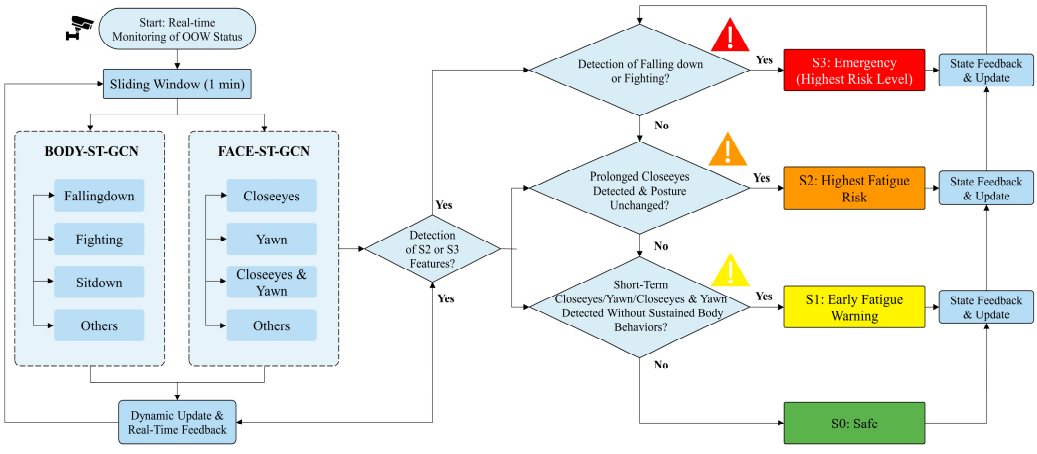


Figure 2. Overall workflow of the dual-branch ST-GCN framework for multimodal behavior detection and hierarchical risk assessment.

3.4. BODY-ST-GCN

3.4.1. GCN with Triple Graph Fusion

Traditional ST-GCN mainly relies on a predefined adjacency matrix to model the physical connections of the human skeleton. While this preserves the prior of skeletal topology, it is insufficient to capture the diverse dependencies among keypoints under complex and varying OOW behaviors. Therefore, a Triple Graph Fusion strategy is proposed and integrated into the GCN module. Specifically, the static graph provides stable topological priors to ensure basic spatial connectivity. The adaptive graph is learned from

data to capture relatively stable correlations beyond physical adjacency. The self-attention graph, generated by an improved multi-head attention module, further models dynamic global dependencies that vary with behaviors. By fusing these three graphs with learnable weights, the model preserves local structural priors while better capturing cross-joint and cross-region interactions, thereby improving the discrimination of complex behavioral patterns. A schematic illustration of the Triple Graph Fusion strategy is shown in Figure 3.

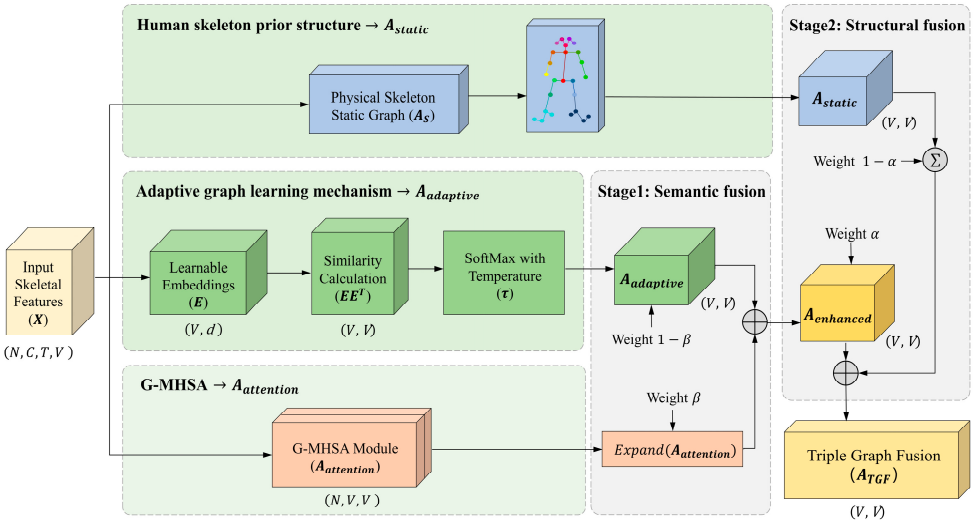


Figure 3. Structure of the TGF strategy.

Specifically, each keypoint in the human skeleton graph is first assigned a learnable embedding vector to form the node embedding matrix E, and each embedding vector in E is normalized to obtain $\hat{E}$. Subsequently, the dot-product similarity between the normalized embedding vectors is computed to construct the semantic similarity matrix S. Finally, the SoftMax function is applied for normalization to generate the adaptive graph $A_{adaptive}$, as shown in Equation (1).

$$A_{adaptive} = softmax\left(\frac{S}{\tau}\right) = softmax\left(\frac{\hat{E}\cdot\hat{E}^T}{\tau}\right) \quad (1)$$

where τ denotes the temperature parameter, which is used to control the smoothness of the $softmax$ function. In this paper, $\tau = 0.1$.

In this paper, a branch for generating spatial attention maps is added to the original MHSA, resulting in the formation of G-MHSA. This branch directly extracts the attention matrices from the MHSA, averages them across multiple heads and temporal steps to construct the adjacency structure of a dynamic graph for graph convolution learning, and produces the final spatial attention map. The detailed computational process of this branch is as follows: for each input feature at the time step t, the Query (Q), Key (K), and Value (V) projection mechanisms are first applied to model the inter among keypoints. Next, a scaled dot-product attention mechanism is employed to compute the spatial similarity between keypoints, yielding the attention matrix $A_h^{(t)}$ for each attention head h at time step t. Finally, within each time step, the attention matrices $A_h^{(t)}$ from different heads are averaged to obtain the spatial attention map of that time step, after which all spatial attention maps across the temporal dimension are further averaged to generate the final spatial attention

map $A_{spatial}$. The computations of $A_h^{(t)}$ and $A_{spatial}$ are shown in Equations (2) and (3), while the structure of the G-MHSA module is shown in Figure 4.

$$A_h^{(t)} = softmax\left(\frac{QK^T}{\sqrt{d_h}}\right) \tag{2}$$

$$A_{spatial} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{H}\sum_{h=1}^{H} A_h^{(t)}\right) \tag{3}$$

where d_h represents the hidden dimension of each attention head; T denotes the total number of time steps; and H is the number of attention heads. In this paper, $H = 8$, while the remaining parameters are automatically determined by the input data or computed within the network.

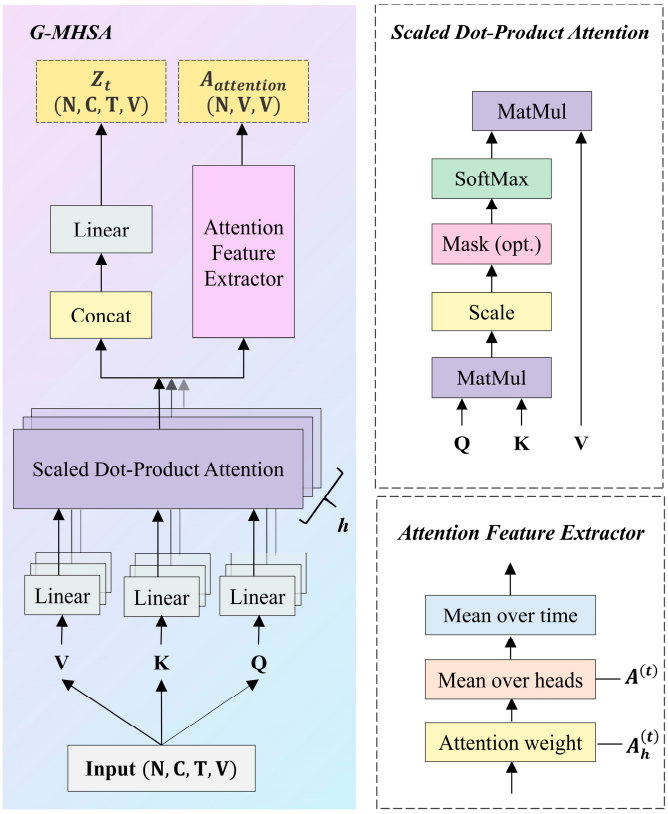


Figure 4. Structure of the G-MHSA module.

Finally, for the network layers incorporating the G-MHSA module, the TGF strategy is applied to obtain the enhanced graph structure. Specifically, the ternary graph fusion strategy first performs semantic-level fusion of $A_{adaptive}$ and $A_{spatial}$ to generate the enhanced adaptive graph $A_{enhanced}$. Subsequently, a structural-level fusion between the skeletal prior static graph A_{static} and $A_{enhanced}$ is conducted to produce the final enhanced graph structure A_{final}. The computations of A_{normal}, $A_{enhanced}$, and A_{final} are provided in Equations (4) and (5).

$$A_{enhanced} = (1-\beta)\cdot A_{adaptive} + \beta \cdot expand\left(A_{spatial}\right) \tag{4}$$

$$A_{final} = (1-\alpha)\cdot A_{static} + \alpha \cdot A_{enhanced} \tag{5}$$

where α and β are defined as learnable scalar parameters. Specifically, α is used to balance the contributions between the static skeleton graph and the adaptive graph, while β is

employed to balance the contributions between the adaptive graph and the attention graph. The raw values of both parameters are initialized to 0.3, and during the forward propagation process, they are mapped to the range of (0, 1) using the sigmoid function, thereby constraining the fusion weights. This constraint ensures that the graph fusion process is consistently formulated as a weighted combination of different graph structures and prevents the occurrence of negative weights or excessively large values that may lead to unreasonable graph representations or unstable model training. Furthermore, both α and β are implemented as globally shared scalar parameters. From a parameterization perspective, this design reduces the complexity of the optimization problem and contributes to improved training stability.

3.4.2. Three-Scale Parallel TCN

In the original ST-GCN, the temporal dimension is modeled using a TCN with fixed convolution kernels. This approach can effectively extract local temporal features between adjacent frames. However, in real-world scenarios, OOW behaviors exhibit significant diversity in the temporal dimension. For example, behaviors such as "fallingdown" are highly abrupt and are mainly characterized by rapid motion changes within a short period of time, whereas behaviors such as "fighting" and "sitdown" are more sustained, and their discriminative information requires modeling the evolution of actions over a longer temporal range. Therefore, temporal convolution with a single receptive field has certain limitations in uniformly modeling complex behaviors. Based on the above analysis, this paper proposes TSP-TCN to replace the conventional single-scale temporal convolution structure in the original ST-GCN.

TSP-TCN improves the temporal convolution structure of the original TCN in ST-GCN. Specifically, the standard convolution in TCN is first replaced with depthwise separable convolution, with the aim of maintaining temporal modeling capability while improving computational efficiency. On this basis, three parallel TCN branches are designed, using convolution kernels of sizes 3×1, 5×1, and 7×1, respectively, to achieve multi-scale temporal modeling. In terms of kernel design, the use of three odd-sized kernels preserves center alignment along the temporal dimension, which is beneficial for stable extraction of temporal features. Meanwhile, the combination of kernel sizes 3, 5, and 7 forms a progressive temporal receptive field from small to large, enabling the model to simultaneously capture both local and broader temporal dependencies. Compared with the original TCN that adopts a single 9×1 kernel, this multi-scale parallel structure provides greater flexibility in modeling the multi-scale dynamic variations in human actions. From the perspective of temporal modeling, the 3×1 kernel mainly focuses on local variations among a few adjacent frames and is more sensitive to fast and subtle motion patterns. The 5×1 kernel models temporal dependencies within a moderate range and can describe the continuous evolution of local actions. The 7×1 kernel provides a wider temporal scope, which helps capture smoother and longer-lasting dynamic trends in actions. It should be noted that the terms short term, medium term, and long term in this work are defined relative to the local temporal receptive field of a single-scale temporal convolution module, rather than the overall duration of an action. Finally, the outputs of the three branches are concatenated along the channel dimension and fed into a fusion layer composed of a 1×1 convolution, a BatchNorm layer, and a ReLU activation function, so as to achieve unified modeling of multi-scale temporal features. The structure of TSP-TCN is shown in Figure 5.

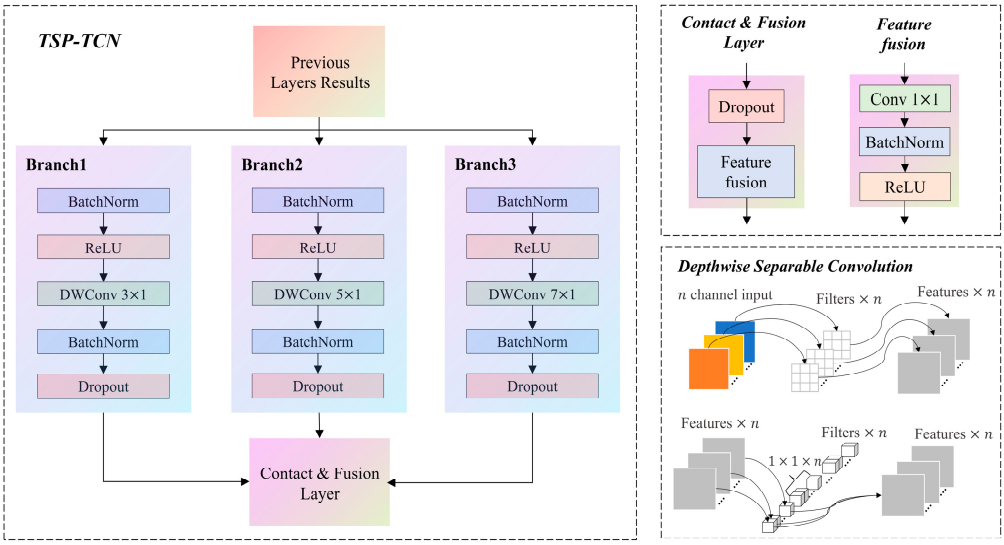


Figure 5. Structure of the TSP-TCN.

Specifically, three parallel temporal feature extraction branches are constructed in TSP-TCN, with convolution kernel sizes of $k_i \times 1$, where $k_i \in \{3, 5, 7\}$. Let the input feature tensor be $X \in R^{C \times T \times V}$. For the i branch, the input feature is first processed through a composite transformation consisting of normalization, an activation function, a temporal convolution with a kernel size $k_i \times 1$, followed by another normalization and dropout, to produce the output Y_i of that branch. The computation process is given in Equation (6).

$$Y_i = \Phi_{k_i}(X),\ i = 1, 2, 3 \tag{6}$$

where $\Phi_{k_i}(\cdot)$ denotes the composite mapping corresponding to the i temporal scale branch.

Subsequently, the outputs of the three branches are concatenated along the channel dimension to obtain the multi-scale temporal feature representation Y_{cat}, as shown in Equation (7).

$$Y_{cat} = Concat(Y_1,\ Y_2,\ Y_3) \tag{7}$$

where the concatenation is performed along the channel dimension, and the resulting feature representation preserves the same dimensional structure as the input tensor.

Finally, a 1×1 convolution is applied to fuse the concatenated features across channels, followed by normalization and activation operations, resulting in the output Y_{out} that integrates multi-scale temporal information. Y_{out} is shown in Equation (8).

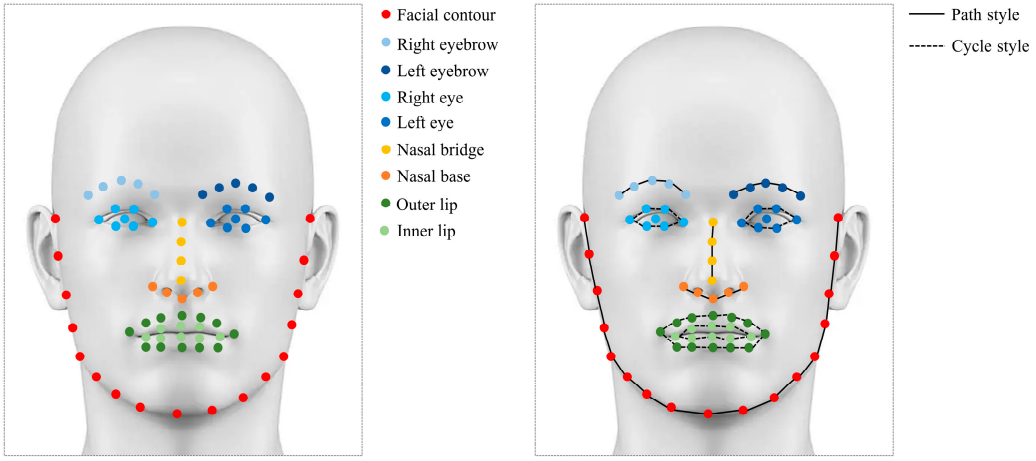
$$Y_{out} = \Psi(Y_{cat}) \tag{8}$$

where $\Psi(\cdot)$ denotes the feature fusion transformation consisting of a 1×1 convolution, normalization, and an activation function. The resulting feature representation preserves the same dimensional structure as the input tensor.

3.5. FACE-ST-GCN

3.5.1. Facial Topology Design

The topology of keypoints defines the adjacency relationship of spatial convolution, which is the basis for ST-GCN to capture the feature behavior of the human body or the face's physical structure in space. However, the facial topology of the OpenPose face model is not defined in ST-GCN, so this paper designs the topology based on the facial keypoint distribution of the face model. The facial topology is shown in Figure 6.



(a) Local region partition (b) Connection method

Figure 6. Illustration of facial landmark topology structure and connection relationships: (**a**) local region partition; (**b**) landmark connection methods (path connections and cycle connections).

The design procedure of the facial topology structure in the face model is as follows. First, a central node is designated to provide a stable spatial reference for the face topology. Since the nose tip is located at the geometric center of the face, the node NoseLower2 at the nose tip is defined as the central node in the facial topology structure of the face model. Next, a self-connection E_{self} is added to each node to ensure that node-specific features can be effectively preserved during the graph convolution operation and to enhance the representation capacity of local structures. Finally, based on the distribution of facial keypoints in the face model, nine local regions are partitioned, and intra-region connections E_{intra} are established within each local region. The representation of the facial topology structure E is shown in Equation (9).

$$E = E_{self} \cup E_{intra} \tag{9}$$

The local regions are divided as follows: facial contour region (FaceContour0–FaceContour16), right eyebrow region (REyeBrow0–REyeBrow4), left eyebrow region (LEyeBrow0–LEyeBrow4), right eye region (REye0–REye5 and RPupil), left eye region (LEye0–LEye5 and LPupil), nasal bridge region (NoseUpper0–NoseUpper3), nasal base region (NoseLower0–NoseLower4), outer lip region (OMouth0–OMouth11), and inner lip region (IMouth0–IMouth7). After the regional partitioning is completed, the intra-region connectivity is defined according to the geometric distribution of the keypoints within each region. For the five relatively open regions in terms of spatial distribution, facial contour, nasal bridge, nasal base, left eyebrow, and right eyebrow, adjacent keypoints are connected pairwise to form a "path-like" connection. For the four relatively closed regions, left eye, right eye, outer lip, and inner lip, an additional connection is introduced between the first and last keypoints on top of the "path-like" connections, thereby forming a "loop-like" connection. Consider the set of keypoints in a region, denoted as $V = \{v_0, v_1, \ldots, v_{n-1}\}$. The sets of keypoint connections defined by the two connection strategies are shown in Equations (10) and (11).

$$E(P_n) = \{(v_i, v_{i+1}) | i = 0,\ 1,\ \ldots, n-2\} \tag{10}$$

$$E(C_n) = \{(v_i, v_{i+1}) | i = 0,\ 1,\ \ldots, n-2\} \cup \{(v_{n-1},\ v_0)\} \tag{11}$$

where $E(P_n)$ denotes the set of keypoint connections formed by the path-like connection strategy; $E(C_n)$ denotes the set of keypoint connections formed by the "loop-like" connection strategy.

3.5.2. TCN with the TAM

Compared with body behaviors, facial fatigue-related behaviors usually exhibit smaller amplitudes, unstable durations, and noticeable individual differences. Therefore, the model is required not only to capture dynamic changes within local temporal ranges but also to effectively model key temporal frames, which places higher demands on its temporal modeling capability. The TAM [37] is an existing temporal modeling method that introduces adaptive weights along the temporal dimension to dynamically adjust features at different time positions, thereby highlighting important information while suppressing redundant information. Based on the characteristic that discriminative information in facial fatigue behaviors is concentrated in specific temporal segments, adopting this module can enhance the model's response to important temporal frames. Specifically, the Temporal Adaptive Module models temporal information by combining a local branch and a global branch. The local branch is designed to capture short-term dynamic changes and generate position-related importance weights. The global branch aggregates global temporal information to generate adaptive temporal convolution kernels, which are used to model temporal dependencies over different time ranges. Compared with conventional temporal convolution networks that use fixed convolution kernels, this module can more flexibly adapt to the uncertainty in the duration of fatigue-related behaviors. The structure of the Temporal Adaptive Module is shown in Figure 7.

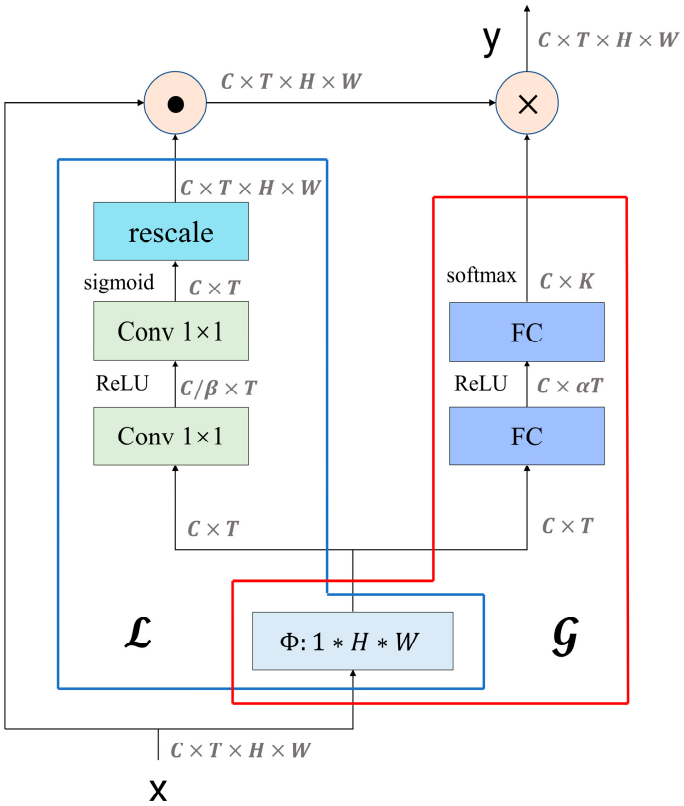


Figure 7. Structure of the TAM.

The local branch is designed to leverage short-term temporal information to generate position-related importance weights. It first learns a position-sensitive importance weight

V to capture the short-term temporal structure, and then perform temporal enhancement. The temporal enhancement process is shown in Equation (12).

$$Z = F_{rescale}(V) \odot X = L(X) \odot X \tag{12}$$

where $\odot$ denotes the element-wise multiplication and $Z \in R^{C \times T \times H \times W}$. To match the size of X, $F_{rescale}(V)$ rescales the V to $\hat{V} \in R^{C \times T \times H \times W}$ by replicating in the spatial dimension.

The global branch is the core of the TAM, as it integrates global contextual information into the TAM and learns position-shared weights for fusion. In response to the diversity of video temporal information, the global branch generates dynamic temporal aggregation convolutional kernels. To simplify the generation of adaptive convolutional kernels, it focuses solely on the temporal modeling aspect when generating the adaptive kernels. The process for generating the convolutional kernels is shown in Equation (13).

$$\Theta_c = G(X)_c = softmax(F(W_2, \delta(F(W_1, \phi(X)_c)))) \tag{13}$$

where $\Theta_c \in R^K$ is generated adaptive kernel (aggregation weights) for c^{th} channel, δ denotes the activation function ReLU.

Finally, the temporal information is adaptively aggregated, and the temporal structural information between video frames is learned through convolution. The entire process is shown in Equation (14).

$$Y_{c,t,j,i} = G(X) \otimes Z = \Theta \otimes Z = \sum_k \Theta_{c,k} \cdot Z_{c,t+k,j,i} \tag{14}$$

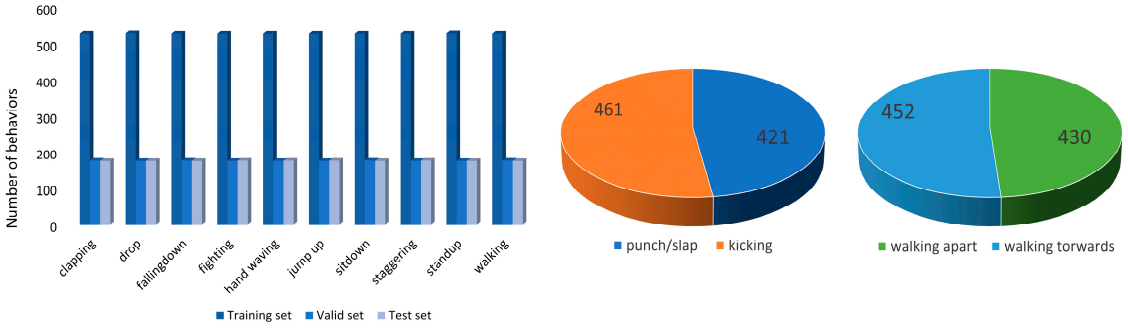
where $\cdot$ denotes the scalar multiplication and Y is the output feature maps ($Y \in \mathbb{R}^{C \times T \times H \times W}$).

4. Experimental Results and Analysis

4.1. Datasets

4.1.1. Construction of a Dataset for Behavior Detection

Compared with land-based tasks, the movements performed by OOW during bridge watchkeeping are relatively simple. Therefore, this paper constructs the 10-Behaviors dataset based on the NTU RGB+D 60 dataset [38] to evaluate the performance of BODY-ST-GCN in human behavior recognition. In this paper, the 10-Behaviors dataset is used to train and evaluate the BODY-ST-GCN model for human action recognition. Since the model relies on human keypoint sequences rather than raw RGB data, it is less sensitive to variations in background and scene conditions. Therefore, the dataset, constructed based on NTU RGB+D 60, is adopted to validate the model's behavior recognition capability. The three unsafe behaviors relevant to OOW watchkeeping examined in this work are also included in this dataset. Specifically, the "fighting" class is composed of selected samples from the "punch/slap" and "kicking" categories of NTU RGB+D 60, while the "walking" class is formed from portions of the "walking towards" and "walking apart" categories. In total, the 10-Behaviors dataset contains 8820 video samples, each lasting 3–5 s with a resolution of 1920 × 1080. The dataset is divided into training, validation, and test sets following a 6:2:2 split ratio. The sample categories and dataset split are illustrated in Figure 8.



(**a**) (**b**)

Figure 8. (**a**) The sample categories and dataset split of the 10-Behaviors dataset. (**b**) Samples of the "fighting" and "walking" behaviors.

4.1.2. Construction of a Dataset for Facial Feature Detection

To construct the facial feature detection dataset, this study builds the Fatigue-Normal dataset based on the publicly available YawDD dataset [39]. The dataset includes four categories of facial states. Among them, "closeeyes", "yawn", and "closeeyes_yawn" are defined as fatigue-related behaviors, while other behaviors without fatigue characteristics, such as normal eye opening and speaking, are collectively categorized as the "normal" state. The Fatigue-Normal dataset consists of 3691 video samples, each with a duration of 3–5 s and a resolution of 640 × 480. The dataset is divided into training, validation, and test sets with a ratio of 6:2:2. The sample categories and dataset split are illustrated in Figure 9.

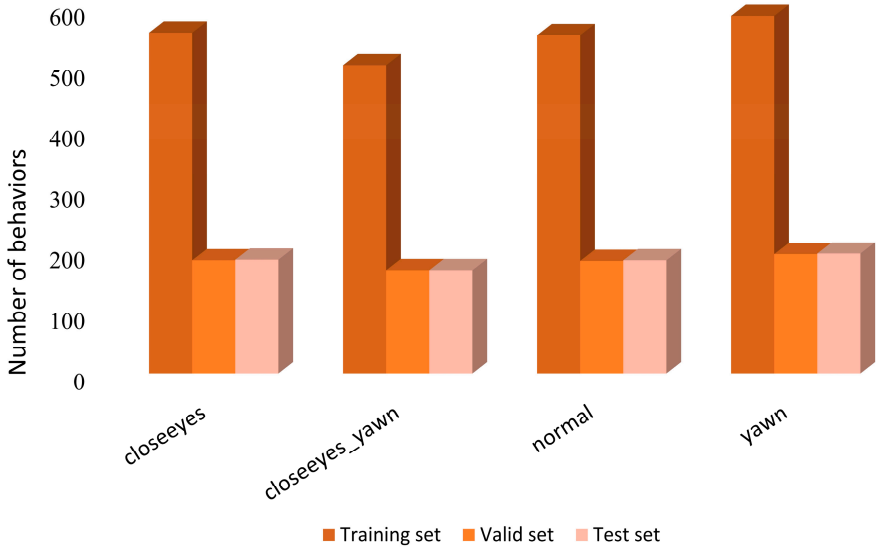


Figure 9. Sample categories and dataset split of the Fatigue-Normal dataset.

4.1.3. Construction of a Realistic Shipboard Dataset

In this paper, an OOW behavior validation dataset was constructed on the bridge of the training vessel "Xin Hongzhuan" at Dalian Maritime University. To simulate OOW behaviors, two male participants, denoted as A and B, were recruited to participate in the experiment. The behavioral data were collected using Hikvision cameras from two viewpoints on the left and right sides of the bridge to enhance viewpoint diversity. The cameras had a resolution of 1920 × 1080 and were installed at a height of 3 m, which is consistent with the standard installation height of surveillance cameras on ship bridges. During the behavior simulation process, participants A and B jointly performed the "fighting"

behavior, while all other OOW duty-related safe and unsafe behaviors were performed solely by participant A. The final self-constructed validation video dataset consists of a total of 22,860 frames. For human behavior annotation, the dataset includes 4530 frames of "fallingdown," 4410 frames of "sitdown," 3630 frames of "fighting," 6780 frames of "standup," and 3510 frames of "walking." In addition, the dataset contains four types of facial behaviors for fatigue detection, including 3930 frames of "closeeyes," 5640 frames of "yawn," 4950 frames of "closeeyes_yawn," and 8340 frames of "normal." The dataset was collected under well-lit daytime watchkeeping conditions and does not involve complex sea states. The data acquisition equipment is shown in Figure 10, and representative samples from the OOW behavior validation dataset are shown in Figure 11. The participant information is shown in Table 2.

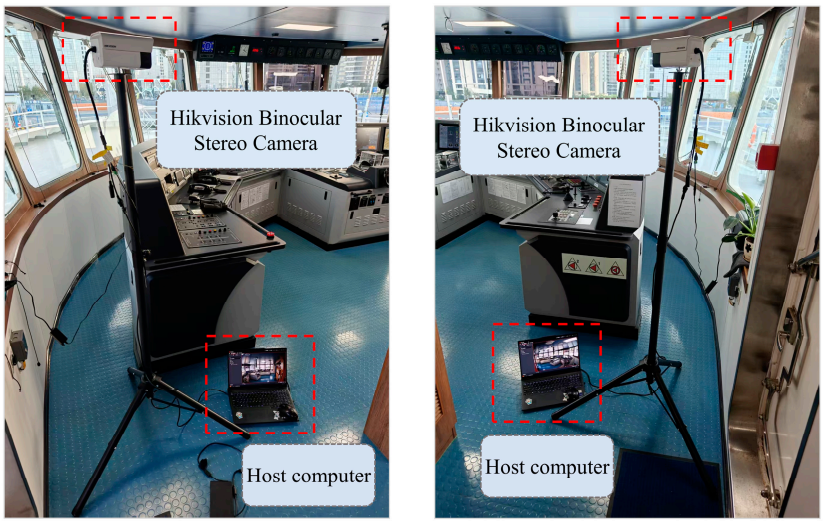


Figure 10. Camera viewpoints and equipment setup on the bridge of the training vessel "Xin Hong Zhuan": (**a**) viewpoint of the left-side camera; (**b**) viewpoint of the right-side camera.



Figure 11. Representative frames of each behavior category in the OOW behavior dataset constructed in a real ship bridge environment: (**a**) standing; (**b**) walking; (**c**) fighting; (**d**) sitting down; (**e**) falling down; (**f**) closing eyes; (**g**) closing eyes and yawning; (**h**) yawning; (**i**) normal.

Table 2. Detailed information about the participants.

Sample	Age	Height	Weight	Sex
Participant A	24	183 cm	86 kg	Male
Participant B	25	179 cm	80 kg	Male

4.2. Experimental Environment and Training Parameters

4.2.1. The BODY-ST-GCN Training Configuration

The training configuration of BODY-ST-GCN is summarized in Table 3. In the experiments on the 10-Behaviors dataset, the model converged and became stable after approximately 200 training iterations; therefore, the total number of training epochs was set to 240. Stochastic Gradient Descent (SGD) was employed as the optimizer, with an initial learning rate of 0.001 and a weight decay of 0.0008. The learning rate was reduced at epochs 120, 180, and 220.

Table 3. Configuration of the BODY-ST-GCN experiment platform.

Configuration	Name	Type
Hardware	GPU	NVIDIA RTX 3060
	Memory	32 GB
Software	Python	3.10
	PyTorch	2.7.1
	CUDA	12.6
Hyperparameters	Learning Rate	0.001
	Optimizer	SGD
	Momentum Value	0.9
	Weight Decay	0.0008
	Batch Size	32
	Training Epochs	240
	Step	[120, 180, 220]

As shown in Figure 12, the validation mean loss of BODY-ST-GCN decreases steadily as the number of epochs increases, with only slight fluctuations observed between epochs 50 and 100 before it quickly resumes its downward trend and stabilizes in the later stages. Similarly, the validation accuracy exhibits a gradual upward trend, experiencing brief fluctuations within the same epoch range before continuing to improve and ultimately remaining at a high level. These results indicate that the model has learned stable and effective feature representations and demonstrates strong predictive performance and generalization capability.

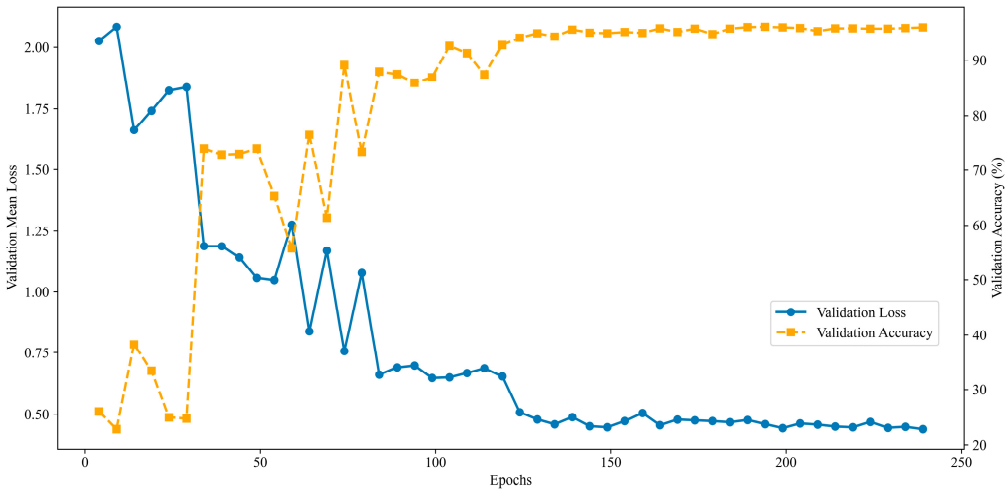


Figure 12. Validation means loss and accuracy of BODY-ST-GCN during the training process.

4.2.2. The FACE-ST-GCN Training Configuration

The training parameters of FACE-ST-GCN are summarized in Table 4. In the experiments on the Fatigue-Normal dataset, the model reached convergence and became stable after approximately 150 training epochs; therefore, the total number of training epochs was set to 200. The optimizer adopted was SGD, with the initial learning rate set to 0.01 and the weight decay set to 0.0001. The learning rate was decayed at the 60th, 100th, and 140th epochs.

Table 4. Configuration of the FACE-ST-GCN experiment platform.

Configuration	Name	Type
Hardware	GPU	NVIDIA RTX 3060
	Memory	32 GB
Software	Python	3.10
	PyTorch	2.7.1
	CUDA	12.6
Hyperparameters	Learning Rate	0.01
	Optimizer	SGD
	Momentum Value	0.9
	Weight Decay	0.0001
	Batch Size	16
	Training Epochs	200
	Step	[60, 100, 140]

Figure 13 shows the variation in validation mean loss and accuracy with respect to epochs for FACE-ST-GCN. As shown in the figure, during the early stages of training, the model exhibits relatively high and fluctuating validation mean loss, indicating that the model has not yet fully learned the data features. As the training progresses, after approximately the 80th epoch, the validation loss gradually stabilizes and remains at a low level, while the accuracy stabilizes at a relatively high range. This behavior suggests that the model has nearly converged and demonstrates good generalization capability on the validation set.

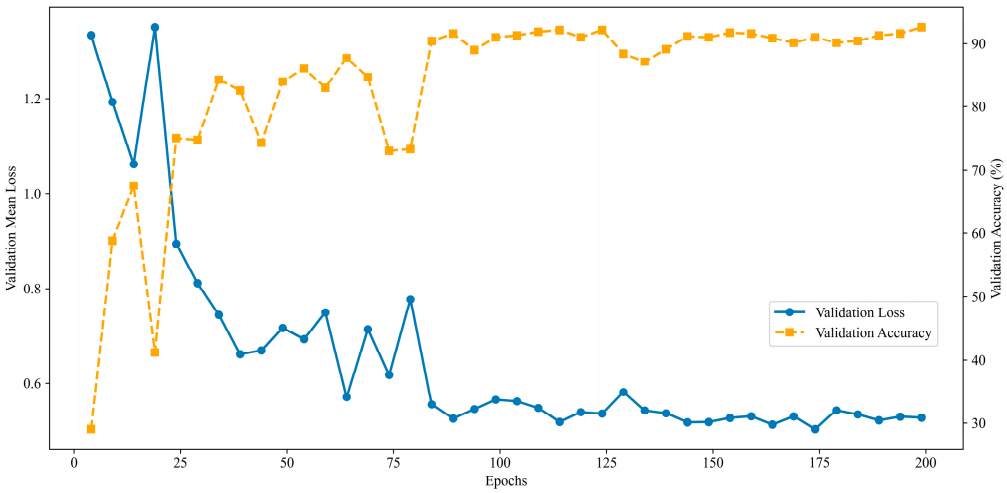


Figure 13. Validation means loss and accuracy of FACE-ST-GCN during the training process.

4.3. Evaluation Metrics

In this paper, Macro-Precision (*Macro-P*), Macro-Recall (*Macro-R*), *Macro-F*1, Giga Floating Point Operations per Second (*GFLOPs*), and Frames Per Second (*FPS*) are adopted

to evaluate the overall performance of the proposed model. *Macro-P*, *Macro-R*, and *Macro-F*1 are defined as the arithmetic means of Precision (*P*), Recall (*R*), and F1-score (*F*1) across all classes, respectively. These metrics provide an overall assessment of the model's classification performance over all behavior categories. Higher values of *Macro-P*, *Macro-R*, and *Macro-F*1 indicate better detection performance across classes. *GFLOPs* represent the number of billions of floating-point operations required for a single forward pass, while *FPS* denotes the number of frames processed per second during inference. A lower *GFLOPs* value indicates a more lightweight model, and a higher *FPS* reflects faster inference speed and better real-time computational efficiency. The formulations of *Macro-P*, *Macro-R*, *Macro-F*1, and *GFLOPs* are shown in Equations (15)–(18).

$$Macro\text{-}P = \frac{1}{k}\sum_{i=1}^{k} P_i \tag{15}$$

$$Macro\text{-}R = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{16}$$

$$Macro\text{-}F1 = \frac{1}{k}\sum_{i=1}^{k} F1_i \tag{17}$$

$$GFLOPs = \frac{Flops}{10^9} \tag{18}$$

where k denotes the number of behavior classes; P_i denotes the precision of class i, which is defined as the ratio of correctly predicted samples of class i to all samples predicted as class i; R_i denotes the recall of class i, which is defined as the ratio of correctly predicted samples of class i to all actual samples of class i; $F1_i$ is the harmonic mean of P_i and R_i; and $Flops$ denotes the total number of floating-point operations required for a single forward pass of the model, which is used to measure the computational complexity of the model.

4.4. Experimental Results of the Sensitivity Analysis

To validate the rationality of the fatigue threshold selection, a sensitivity analysis was conducted on three PERCLOS thresholds, and four sensitivity evaluation metrics were adopted for assessment. The results are shown in Figure 14. Specifically, Figure 14a presents the fatigue window count, which represents the total number of time windows classified as fatigue and reflects the overall duration of fatigue states. Figure 14b shows the fatigue window ratio, defined as the proportion of fatigue windows among all time windows, which measures the overall level of fatigue. Figure 14c illustrates the fatigue episode count, representing the number of consecutive fatigue window sequences and describing the frequency of fatigue occurrence. Figure 14d shows the average episode length, which characterizes the average duration of a single fatigue event. As observed from the figure, when the threshold increases from 50% to 70%, the fatigue window count, fatigue window ratio, and fatigue episode count decrease significantly, while the average episode length increases notably, indicating that the model is sensitive to threshold variations within this range. When the threshold further increases from 70% to 80%, all evaluation metrics remain relatively stable, suggesting that the model becomes insensitive to threshold changes and exhibits strong stability in this interval. Therefore, the range of 70% to 80% can be regarded as a stable interval for eye-based fatigue determination. Furthermore, considering that a higher threshold may impose overly strict criteria and lead to missed detections of fatigue states in practical applications, the lower bound of this stable interval, namely 70%, is selected as the final fatigue determination threshold. This choice ensures the stability of detection results while maintaining a balance between sensitivity and reliability.

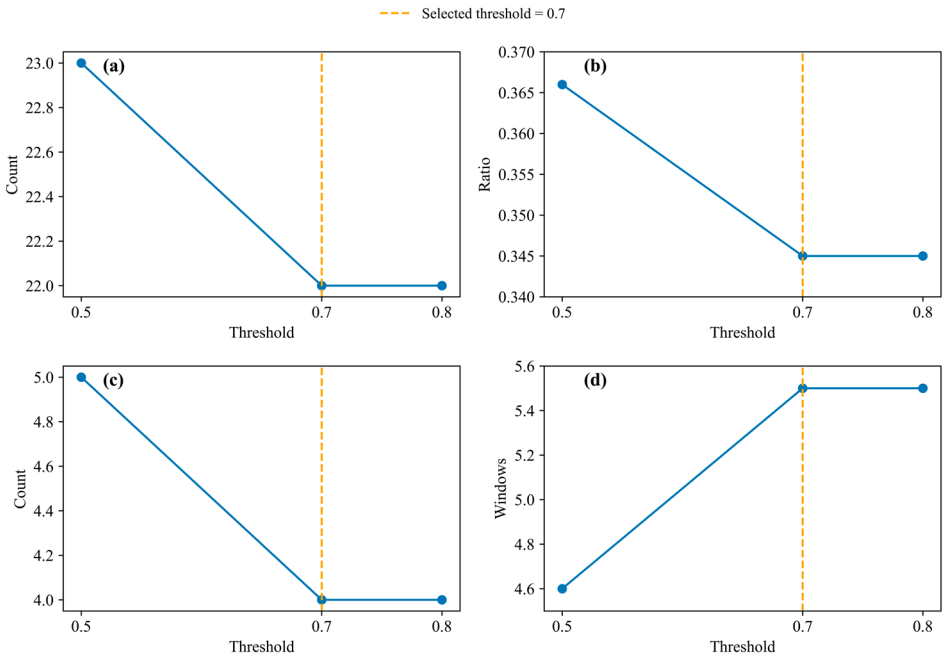


Figure 14. Sensitivity analysis of fatigue detection results under different PERCLOS thresholds: (**a**) fatigue window count, (**b**) fatigue window ratio, (**c**) fatigue episode count, and (**d**) average episode length.

To analyze the impact of the temperature parameter τ on model performance, sensitivity analysis experiments are conducted on both BODY-ST-GCN and FACE-ST-GCN. In the experiments, τ is set within the range of [0.05, 0.10, 0.15, 0.20, 0.25], and *Macro-F*1, which reflects the overall performance of the model, is used as the evaluation metric. The experimental results are shown in Figure 15. It can be observed that under different values of τ, the *Macro-F*1 scores of both models vary only slightly, indicating a certain level of stability. Among them, when τ is set to 0.1, the models achieve relatively better performance, with *Macro-F*1 scores of 0.969 and 0.938 for BODY-ST-GCN and FACE-ST-GCN, respectively. Therefore, τ is set to 0.1 in the subsequent experiments.

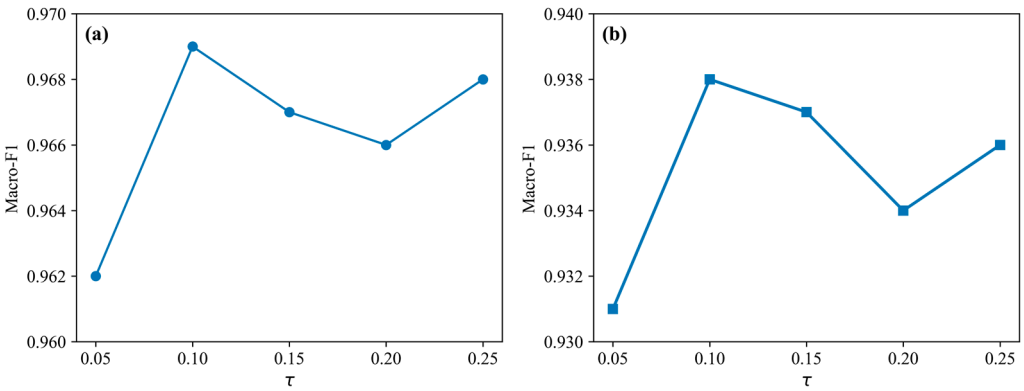


Figure 15. Sensitivity analysis of the temperature parameter τ: (**a**) effect of τ on the performance of BODY-ST-GCN; (**b**) effect of τ on the performance of FACE-ST-GCN.

To analyze the impact of the number of attention heads H in G-MHSA on model performance, sensitivity analysis experiments are conducted on both BODY-ST-GCN and FACE-ST-GCN. In the experiments, H is set to [2, 4, 8, 16], and *Macro-F*1 is again adopted as the evaluation metric for analysis. The experimental results are shown in Figure 16. As can be seen from Figure 16a, in BODY-ST-GCN, as H increases from two to eight, the *Macro-F*1 score shows a continuous upward trend; when H is further increased to 16, the *Macro-F*1

remains the same as that at $H = 8$. This indicates that the performance of BODY-ST-GCN tends to stabilize when $H = 8$. From Figure 16b, it can be observed that as H increases from two to eight, the *Macro-F*1 of FACE-ST-GCN also shows an upward trend; however, when H is increased to 16, the *Macro-F*1 decreases slightly. FACE-ST-GCN still achieves the best *Macro-F*1 when $H = 8$. Therefore, in the subsequent experiments, the number of attention heads H in G-MHSA is set to eight.

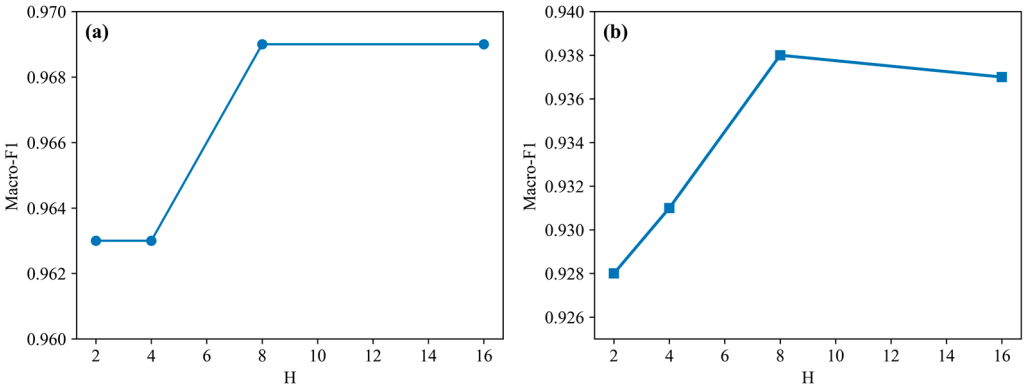


Figure 16. Sensitivity analysis of the number of attention heads (H) in G-MHSA. (**a**) Macro-F1 performance of BODY-ST-GCN across different H settings. (**b**) Macro-F1 performance of FACE-ST-GCN across different H settings.

4.5. Experimental Results of the Ablation

4.5.1. Ablation Experiment for Different Graph Structures

To evaluate the contribution of different graph structures in the TGF strategy, three variant models are constructed based on ST-GCN, including:

- ADA-ST-GCN: An adaptive graph learning mechanism is introduced into the baseline ST-GCN to evaluate the impact of the adaptive graph structure on model performance.
- G-MHSA-ST-GCN: A G-MHSA module is incorporated into the baseline ST-GCN to investigate the effect of the attention-based graph structure on model performance.
- TGF-ST-GCN: A TGF strategy is introduced in the ST-GCN, and through a comparison with the baseline model, the effectiveness of the TGF strategy in improving the model's performance is validated.

Ablation experiments are conducted on the 10-Behaviors and Fatigue-Normal datasets to analyze the impact of different graph structures on the discriminative ability of the two models. Since this part mainly focuses on model discrimination performance, and the differences in computational complexity and inference efficiency among the variants are relatively small, *Macro-P*, *Macro-R*, and *Macro-F*1 are adopted as the evaluation metrics.

Table 5 presents the experimental results on the 10-Behaviors dataset. Compared with the baseline model ST-GCN, ADA-ST-GCN with the adaptive graph achieves consistent improvements across all evaluation metrics. With the further introduction of G-MHSA, the model performance is continuously improved, indicating that the attention mechanism can effectively capture global dependencies among keypoints. Finally, TGF-ST-GCN, which fuses the three types of graph structures, achieves the best performance, improving the *Macro-F*1 score to 0.944. These results demonstrate that different graph structures are complementary in spatial modeling, and their fusion can significantly enhance the model's ability to recognize complex behaviors.

Table 5. Ablation experimental results of different graph structures on the 10-Behaviors dataset.

Module	*Macro-P*	*Macro-R*	*Macro-F*1
ST-GCN	0.910	0.908	0.907
ADA-ST-GCN	0.921	0.910	0.910
G-MHSA-ST-GCN	0.934	0.927	0.927
TGF-ST-GCN	**0.945**	**0.943**	**0.944**

Note: The bold black font is the best value.

Table 6 presents the experimental results on the Fatigue-Normal dataset. Overall, the introduction of different graph structures consistently improves model performance. Compared with the baseline ST-GCN, ADA-ST-GCN achieves moderate gains across all metrics, indicating that the adaptive graph enhances the modeling of relationships among facial keypoints. G-MHSA-ST-GCN further improves performance, especially in *Macro-R* and *Macro-F*1, demonstrating the effectiveness of modeling global dependencies for capturing subtle fatigue-related features. Finally, TGF-ST-GCN achieves the best performance across all metrics, with *Macro-F*1 reaching 0.905. This indicates that the fusion of multiple graph structures can more effectively model complex facial relationships and improve fatigue recognition performance.

Table 6. Ablation experimental results of different graph structures on the Fatigue-Normal dataset.

Module	*Macro-P*	*Macro-R*	*Macro-F*1
ST-GCN	0.822	0.821	0.820
ADA-ST-GCN	0.879	0.824	0.825
G-MHSA-ST-GCN	0.908	0.880	0.879
TGF-ST-GCN	**0.918**	**0.902**	**0.905**

Note: The bold black font is the best value.

4.5.2. Ablation Experiment for BODY-ST-GCN

In this section, ablation experiments are designed to demonstrate the effectiveness of BODY-ST-GCN. Building upon the graph structure variants introduced above, four models are constructed based on different improvement strategies to verify the necessity of each component. The baseline ST-GCN is first considered as a reference model without any modifications. In addition, the TGF-ST-GCN, which has been described in the previous subsection, is included to further evaluate the contribution of graph structure fusion within the overall framework. Furthermore, a TSP-ST-GCN model is constructed by replacing the original TCN with the proposed TSP-TCN, aiming to verify the effectiveness of multi-scale temporal modeling. Finally, the proposed BODY-ST-GCN integrates both the TGF strategy in the GCN component and the TSP-TCN in the temporal component, so as to simultaneously enhance spatial and temporal modeling capabilities and achieve better performance in detecting OOW behaviors.

To clearly assess the learning and discrimination capability of each model, four ablation variants were evaluated on the test set of the 10-Behaviors dataset, and the high-dimensional features of all behavior classes were visualized using t-SNE. Figure 17 presents the t-SNE feature distributions for each model: (a) baseline ST-GCN, (b) TGF-ST-GCN, (c) TSP-ST-GCN, and (d) BODY-ST-GCN. As shown in Figure 17a, the baseline ST-GCN forms clusters for most classes but exhibits loose intra-class distributions and a considerable number of outliers, particularly for “drop”, “fallingdown”, and “handwaving”. With enhanced spatial modeling, TGF-ST-GCN improves the compactness of most class clusters. TSP-ST-GCN further strengthens temporal modeling and yields clearer class boundaries and fewer outliers compared with ST-GCN. Combining both spatial and temporal enhancement strategies, BODY-ST-GCN achieves the most compact intra-class clustering and the

most distinct inter-class separation, with the fewest outliers across all behavior categories, indicating a more stable and discriminative feature space.

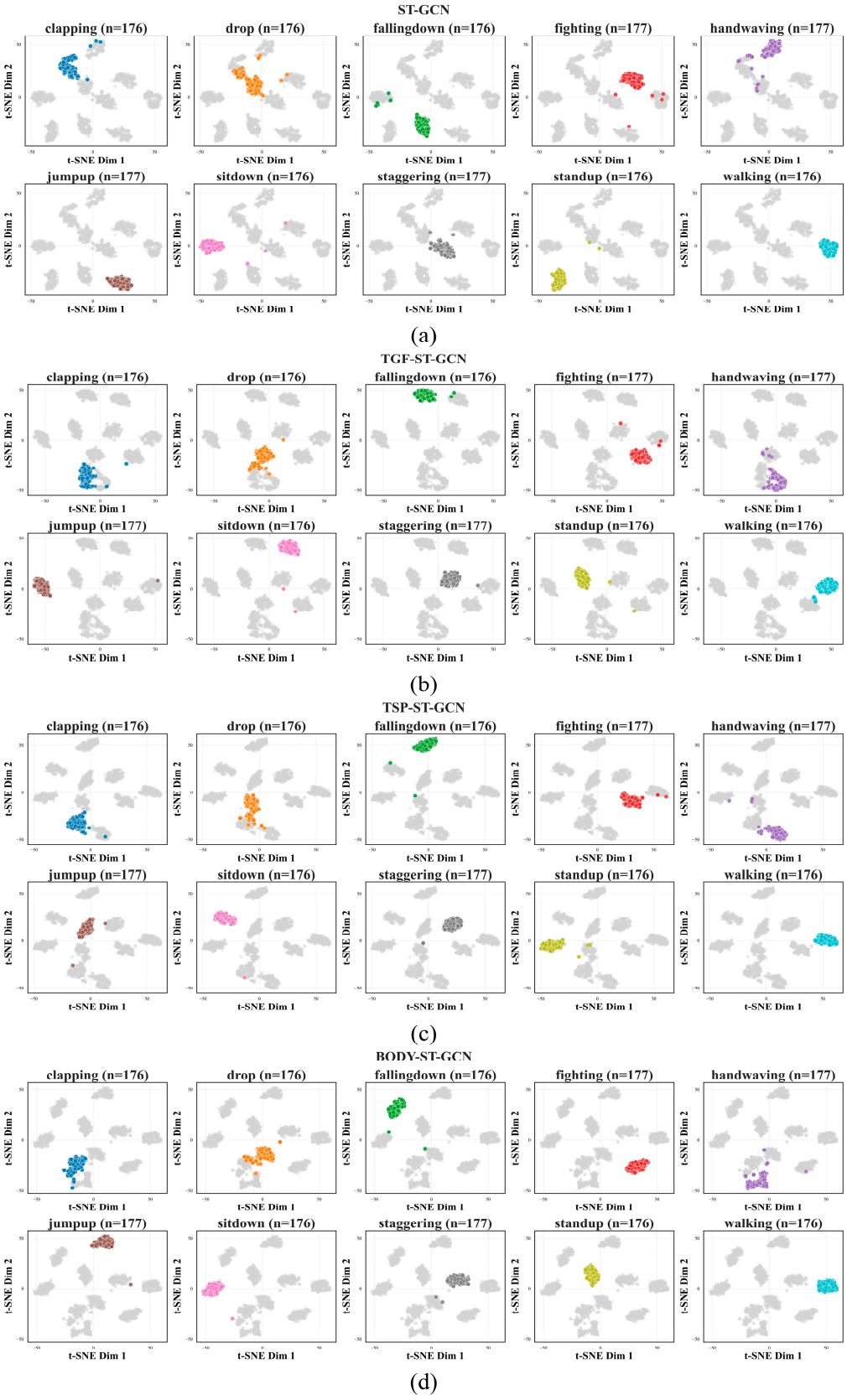


Figure 17. Behavior recognition for different ablation models with t-SNE visualization: (**a**) ST-GCN; (**b**) TGF-ST-GCN; (**c**) TSP-ST-GCN; and (**d**) BODY-ST-GCN.

To further analyze the discriminative ability of each model, misclassified samples in the test set were marked with red crosses, as shown in Figure 18. Figure 18a–d correspond to ST-GCN, TGF-ST-GCN, TSP-ST-GCN, and BODY-ST-GCN, respectively. The baseline ST-GCN misclassified 81 samples, mainly from the "drop", "fighting", "handwaving", "sitdown", "staggering", and "walking" classes. In contrast, all enhanced variants significantly reduced the number of misclassified samples. Benefiting from improved spatial and temporal modeling, TGF-ST-GCN and TSP-ST-GCN reduced the misclassification counts to 66 and 59, respectively. Finally, although BODY-ST-GCN still shows room for improvement in distinguishing walking and dropping, it achieves the lowest number of misclassified samples among all models, further confirming its superior discriminative capability across behavior classes.

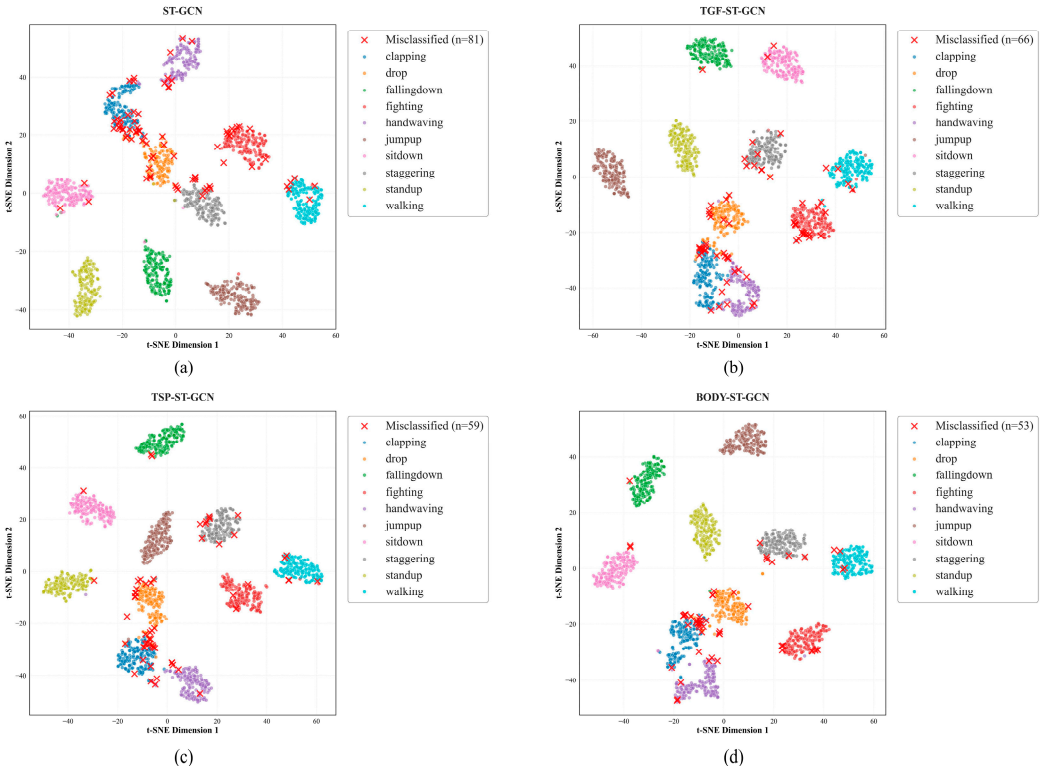


Figure 18. Misclassified samples recognition for different ablation models with t-SNE visualization: (**a**) ST-GCN; (**b**) TGF-ST-GCN; (**c**) TSP-ST-GCN; and (**d**) BODY-ST-GCN.

Table 7 presents the performance of the ablation models on the test set of the 10-Behaviors dataset. The results show that all models incorporating the proposed enhancements outperform the baseline ST-GCN to varying degrees. Specifically, BODY-ST-GCN, which integrates both the TGF and TSP strategies and strengthens spatial and temporal modeling simultaneously, achieves the highest *Macro-P*, *Macro-R*, and *Macro-F*1 values. Compared with ST-GCN, these three metrics improve by 6.4%, 6.6%, and 6.8%, respectively. Likewise, TGF-ST-GCN and TSP-ST-GCN, which introduce only spatial or temporal enhancement, also achieve noticeable improvements over the baseline.

From the perspective of computational efficiency, ST-GCN exhibits the lowest $GFLOPs$, yet its FPS is lower than those of the enhanced models. TGF-ST-GCN attains the highest FPS, primarily because TGF replaces the sparse matrix multiplication used in ST-GCN with adaptive adjacency learning and trainable node embeddings, leading to more regularized computations during inference. In addition, the G-MHSA module in TGF performs averaging in the temporal dimension, further reducing computational cost. The TSP strategy replaces the original large-kernel TCN in ST-GCN with parallel small-kernel branches,

which not only lowers the computational burden but also improves GPU utilization, resulting in higher *FPS* despite slightly increased *GFLOPs*. Although BODY-ST-GCN has a lower *FPS* than TGF-ST-GCN, it consistently achieves superior macro-level performance across all evaluation metrics.

Table 7. Ablation experimental results of model accuracy on the 10-Behaviors dataset.

Module	*Macro-P*	*Macro-R*	*Macro-F*1	*GFLOPs*	*FPS*
ST-GCN	0.910	0.908	0.907	**5.875**	46.002
TGF-ST-GCN	0.945	0.943	0.944	5.876	**84.718**
TSP-ST-GCN	0.950	0.949	0.949	7.410	67.107
BODY-ST-GCN	**0.969**	**0.968**	**0.969**	7.411	61.503

Note: The bold black font is the best value.

Figure 19 presents the confusion matrices on the 10-Behaviors dataset. As shown, the baseline ST-GCN achieves relatively high recognition rates for behaviors such as "sitdown", "standup", and "walking", but exhibits considerable confusion among several classes. For example, 9.66% and 8.52% of "drop" samples are misclassified as "clapping" and "fallingdown", while 1.13%, 2.26%, and 6.21% of "jumpup" samples are misclassified as "drop", "handwaving", and "standup", respectively. With enhanced spatial modeling, TGF-ST-GCN significantly reduces misclassification across most behavior categories compared with the baseline. The TSP strategy enables the model to better adapt to actions with different temporal scales, leading to marked performance improvements for medium-duration behaviors such as "fallingdown" and "sitdown", as well as short-duration behaviors such as "fighting", demonstrating the effectiveness of TSP in strengthening temporal modeling. By jointly enhancing spatial and temporal modeling, BODY-ST-GCN achieves the most accurate and balanced classification performance. For the five key behaviors emphasized in this paper, namely "fallingdown", "fighting", "sitdown", "standup", and "walking", BODY-ST-GCN attains classification accuracies exceeding 98%, consistently outperforming the models that incorporate only a single enhancement strategy. Although its accuracy for "sitdown" is slightly lower than that of the baseline, BODY-ST-GCN demonstrates superior and more stable recognition capability across all other behaviors, resulting in overall performance that far surpasses the baseline.

The complete BODY-ST-GCN architecture is a hierarchical model composed of ten ST-GCN blocks. To more precisely determine the optimal insertion position of the TGF module, we construct a set of model variants that differ only in the location where TGF is integrated. The backbone can be conceptually divided into three stages. Blocks 1–4 form the shallow stage with 64 channels and are primarily responsible for extracting low-level action semantics. After the first spatiotemporal downsampling, Blocks 5–7 constitute the intermediate stage with 128 channels, where local semantic relationships are progressively aggregated. Following the second spatiotemporal downsampling, Blocks 8–10 comprise the deep stage with 256 channels, capturing high-level and global action semantics. Based on this hierarchical structure, six representative insertion positions are selected: Blocks 2 and 4 in the shallow stage, Blocks 5 and 7 in the intermediate stage, and Blocks 8 and 10 in the deep stage. These positions span the progression from low-level to high-level semantic learning and cover both spatiotemporal downsampling transitions within the network.

Table 8 summarizes the performance of the models with TGF inserted at different locations. As shown, the insertion position produces negligible changes in computational complexity, and all model variants sustain inference speeds of over 60 *FPS*, indicating that real-time capability remains largely unaffected. However, the recognition accuracy varies markedly with the insertion position. The best *Macro-P*, *Macro-R*, and *Macro-F*1 are achieved when TGF is inserted at Block 4, followed by insertion at Block 10. In contrast,

inserting TGF too early, for example, at Block 2, or placing it within the intermediate stage, leads to noticeable performance degradation. In summary, Block 4 is selected as the optimal insertion position for TGF. Although this configuration does not yield the highest *FPS*, it provides the best overall recognition accuracy, making it the most effective choice for the proposed architecture.

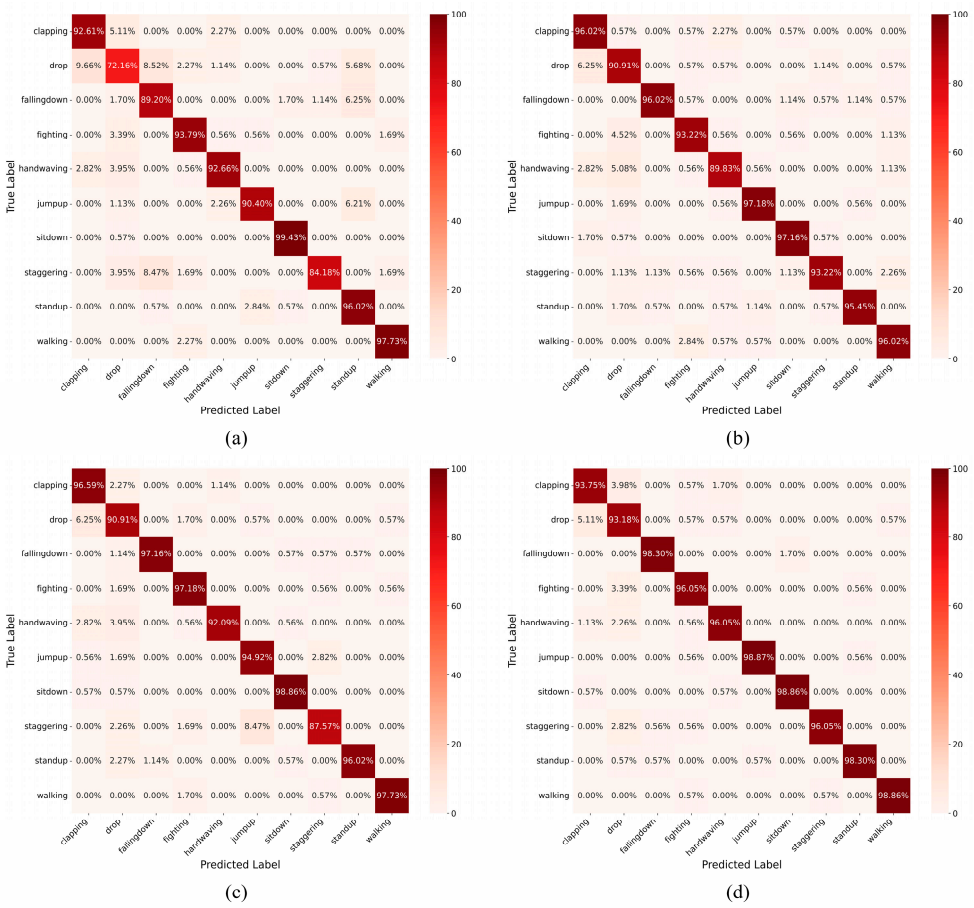


Figure 19. Comparison of confusion matrices for the four ablation models on the 10-Behaviors dataset: (**a**) ST-GCN; (**b**) TGF-ST-GCN; (**c**) TSP-ST-GCN; and (**d**) BODY-ST-GCN.

Table 8. The impact of TGF insertion location on model performance.

Block	*Macro-P*	*Macro-R*	*Macro-F1*	*GFLOPs*	*FPS*
None (baseline)	0.950	0.949	0.949	7.410	67.107
Block 2	0.949	0.947	0.947	**7.410**	60.424
Block 5	0.959	0.958	0.958	7.411	59.679
Block 7	0.947	0.945	0.945	**7.410**	60.722
Block 8	0.953	0.951	0.952	**7.410**	61.760
Block 10	0.966	0.965	0.966	7.411	**64.528**
Ours	**0.969**	**0.968**	**0.969**	7.412	61.503

Note: The bold black font is the best value.

To evaluate the impact of convolutional kernel scales on temporal feature modeling, we take the single-scale convolution (kernel size = 9) used in ST-GCN as the baseline and construct several TCN variants with different temporal scales and kernel sizes within the BODY-ST-GCN framework. First, under the single-scale setting, we set the kernel sizes to one and five to examine the effect of short and medium temporal windows on model performance. Then, we further design dual-scale (1, 5) and tri-scale (1, 3, 5) structures to simultaneously capture temporal dependencies at different scales. Finally, these configura-

tions are compared with the proposed TSP structure (3, 5, 7) to comprehensively assess its effectiveness.

The ablation results are summarized in Table 9. They show that, although the single-scale models with short and medium temporal windows achieve higher real-time performance, their *Macro-P*, *Macro-R*, and *Macro-F*1 scores are significantly lower than those of the multi-scale models. The dual-scale model yields a certain performance gain, but still lags behind the baseline model with a long temporal window in terms of both accuracy and real-time performance. In contrast, the proposed TSP structure maintains competitive real-time performance while improving *Macro-P*, *Macro-R*, and *Macro-F*1 by 2.5%, 2.6%, and 2.6%, respectively, over the baseline single-scale long-window model. Moreover, the TSP-based model exhibits markedly better real-time performance than the alternative tri-scale (1, 3, 5) configuration.

Table 9. The impact of different scale kernels on model performance.

Scale	Kernel Size	*Macro-P*	*Macro-R*	*Macro-F*1	*GFLOPs*	*FPS*
Single	(1)	0.774	0.765	0.765	**1.983**	**121.296**
	(5)	0.922	0.919	0.920	3.929	103.344
	(9)	0.945	0.943	0.944	**5.876**	84.718
Double	(1, 5)	0.935	0.933	0.934	4.460	62.474
Triple	(1, 3, 5)	0.963	0.961	0.962	10.330	38.445
	Ours	**0.969**	**0.968**	**0.969**	7.411	61.503

Note: The bold black font is the best value.

Figure 20a–e illustrate the detection results of various ablation models during the process of a person transitioning from standing to falling. As shown between Figure 20b and Figure 20c, the person exhibits motion blur, with their posture resembling the "sitdown" action. The baseline model, ST-GCN, misclassifies this phase as "sitdown," only correctly identifying the "fallingdown" behavior between Figure 20d and Figure 20e, when the individual assumes a complete falling posture. In contrast, TGF-ST-GCN, TSP-ST-GCN, and BODY-ST-GCN avoid this misclassification. Notably, BODY-ST-GCN recognizes the person's behavior as "staggering" in Figure 20b, demonstrating its ability to capture subtle movements, and continues to accurately detect "fallingdown" in the subsequent frames. Moreover, the confidence score gradually increases from 0.894 to 0.952 as the falling posture becomes more stable, reaching the highest value among all compared models.

Figure 21a–e present the change detection results of various ablation models during the transition from "standup" to "sitdown". In Figure 21a, all models detect the person's behavior as "standup." However, as time progresses and the person's posture changes, the baseline model, ST-GCN, incorrectly classifies the "sitdown" behavior as "jumpup" in Figure 21c. In contrast, other models, after incorporating different improvement strategies, successfully detect this transitional phase. As the person remains seated, while TGF-ST-GCN and TSP-ST-GCN maintain high detection confidence, they exhibit fluctuations due to minor changes in the seated posture. Meanwhile, BODY-ST-GCN not only avoids misclassification of the subject's behavior but also achieves consistently high and stable confidence in recognizing the "sitdown" action. As shown in Figure 21d,e, when the seated posture becomes stable, the confidence scores for "sitdown" reach 0.934 and 0.941, respectively, which are the highest among all ablation models.

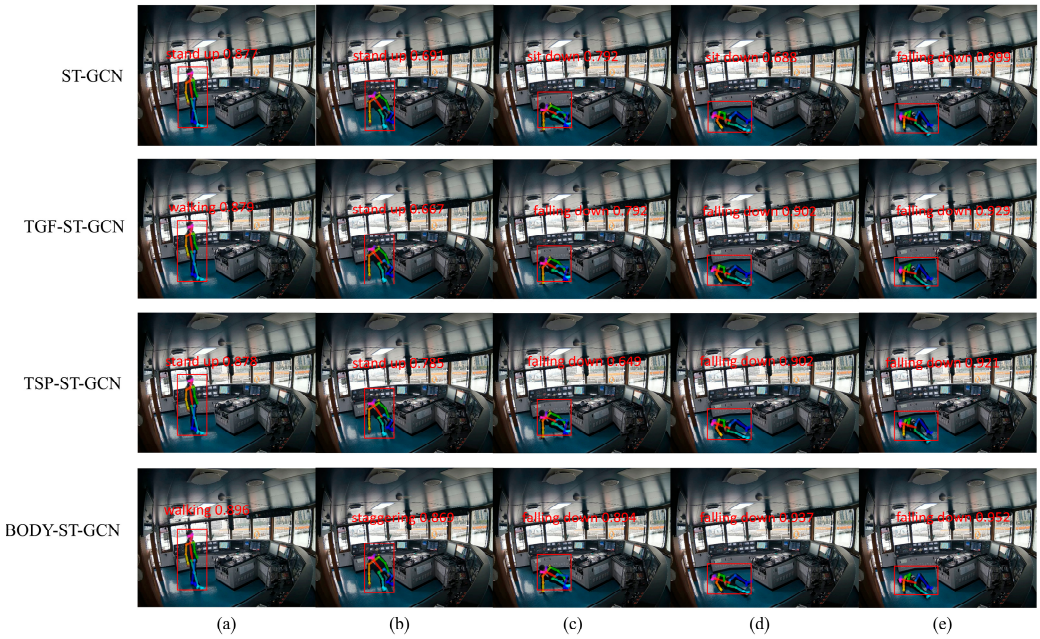


Figure 20. Comparison of four ablation models in detecting the temporal evolution of the "falling-down" behavior: (**a**) detection results of the "standup" behavior; (**b**) detection results of the "staggering" behavior; (**c**) detection results for the initial phase of the "fallingdown" behavior; (**d**) detection results for the execution phase of the "fallingdown" behavior; (**e**) detection results for the completion phase of the "fallingdown" behavior.

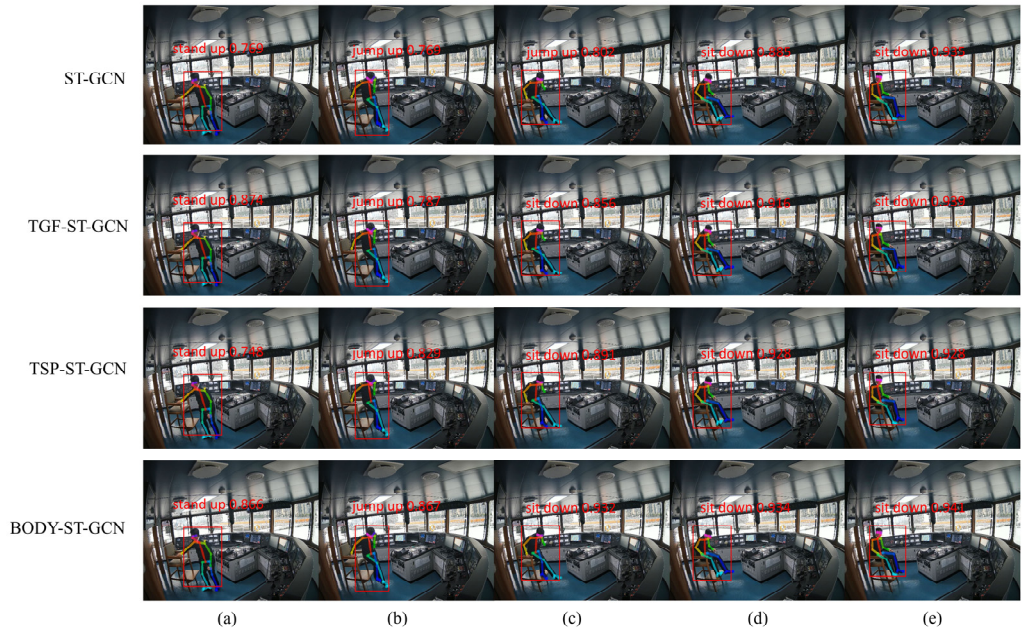


Figure 21. Comparison of four ablation models in detecting the temporal evolution of the "sitdown" behavior: (**a**) detection results of the "standup" behavior; (**b**) detection results of the "jumpup" behavior; (**c**) detection results for the initial phase of the "sitdown" behavior; (**d**) detection results for the execution phase of the "sitdown" behavior; (**e**) detection results for the completion phase of the "sitdown" behavior.

Figure 22a–e show the performance of various ablation models in detecting the "fighting" behavior. "Fighting" is a relatively complex interpersonal interaction, with most body movements concentrated in the arm region. As a result, the baseline model, ST-GCN, initially fails to detect this behavior accurately, misclassifying it as "standup." Only as the person's posture evolves does ST-GCN begin to correctly detect the "fighting" behavior. In

contrast, TGF-ST-GCN accurately detects the "fighting" behavior in most frames, although it makes a misclassification in Figure 22c. TSP-ST-GCN captures more comprehensive features over longer time scales, but due to its limited spatial feature modeling ability, similar to ST-GCN, it misclassifies the behavior as "walking" in Figure 22a, before correctly identifying "fighting" in subsequent frames. BODY-ST-GCN, on the other hand, consistently detects the "fighting" behavior accurately across all frames. Although its detection confidence is lower in Figure 22a,b, it maintains a high confidence level of above 0.85 in subsequent frames. Although the detection performance of all models improves in the later frames, the key differences are mainly reflected in the early-stage recognition capability and confidence levels. Therefore, BODY-ST-GCN demonstrates superior robustness and discriminative ability in handling complex interactive behaviors such as "fighting".

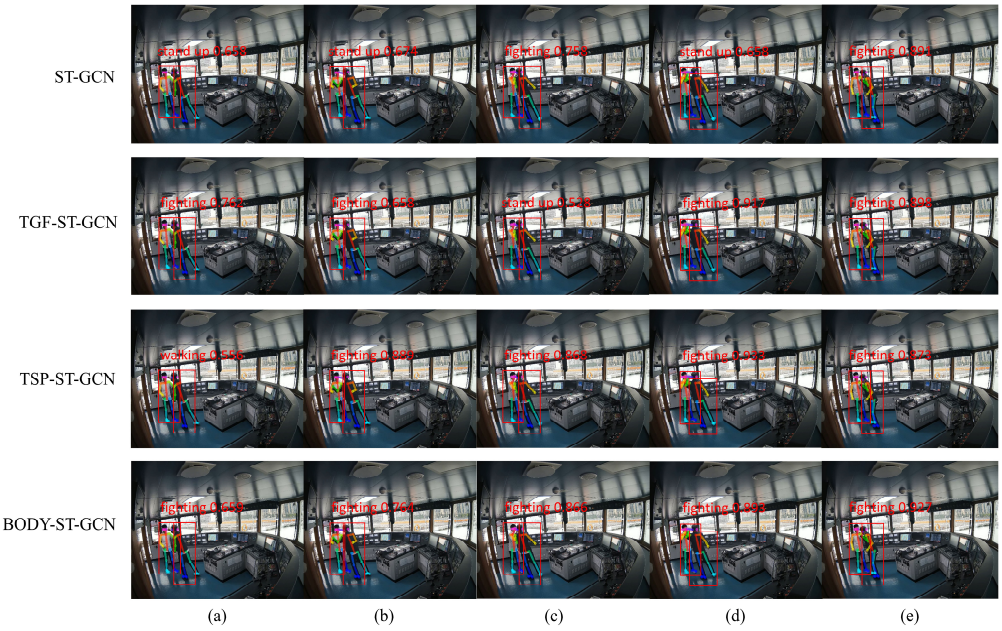


Figure 22. Comparison of four ablation models in detecting the "fighting" behavior across different phases: (**a**) detection results for the "fighting" behavior; (**b**) detection results for the "fighting" behavior; (**c**) detection results for the "fighting" behavior; (**d**) detection results for the "fighting" behavior; (**e**) detection results for the "fighting" behavior.

4.5.3. Ablation Experiments for FACE-ST-GCN

In this section, ablation experiments are designed to demonstrate the effectiveness of FACE-ST-GCN. Similarly to the ablation experiments conducted for BODY-ST-GCN, four deep learning models are constructed based on the different improvement strategies incorporated into FACE-ST-GCN, in order to validate the effectiveness of each strategy. ST-GCN is used as the baseline model without any improvement strategies, serving as a reference for performance comparison with the subsequent improved models. In addition, the TGF-ST-GCN, as described in the previous subsection, is included to evaluate the contribution of the TGF strategy within the facial feature recognition framework. Furthermore, a TAM-ST-GCN model is constructed by incorporating the TAM solely into the TCN component of the baseline model, aiming to assess whether the introduction of TAM can enhance the detection capability for different facial features. Finally, the proposed FACE-ST-GCN integrates both the TGF strategy and the TAM into the baseline model. By combining the advantages of these two strategies, this model is designed to enhance spatiotemporal modeling capabilities and further improve recognition performance for facial features.

Figure 23 shows the t-SNE visualizations of high-dimensional behavior features extracted by the four models on the test set of the Face-Normal dataset. Specifically, Figure 23a corresponds to the baseline ST-GCN, Figure 23b to TGF-ST-GCN, Figure 23c to TAM-ST-GCN, and Figure 23d to the proposed FACE-ST-GCN. As observed, the feature distributions of the four behavior classes in ST-GCN are the most dispersed, indicating limited discriminative capability. In contrast, both TGF-ST-GCN and TAM-ST-GCN exhibit improved feature compactness, with more evident clustering, particularly for the "closeeyes_yawn" or "closeeyes" classes. For FACE-ST-GCN, although a small number of outliers remain, the intra-class compactness is significantly enhanced compared with the other models, and the clusters corresponding to the "closeeyes_yawn" and "yawn" behaviors are more regular in shape with clearer inter-class boundaries.

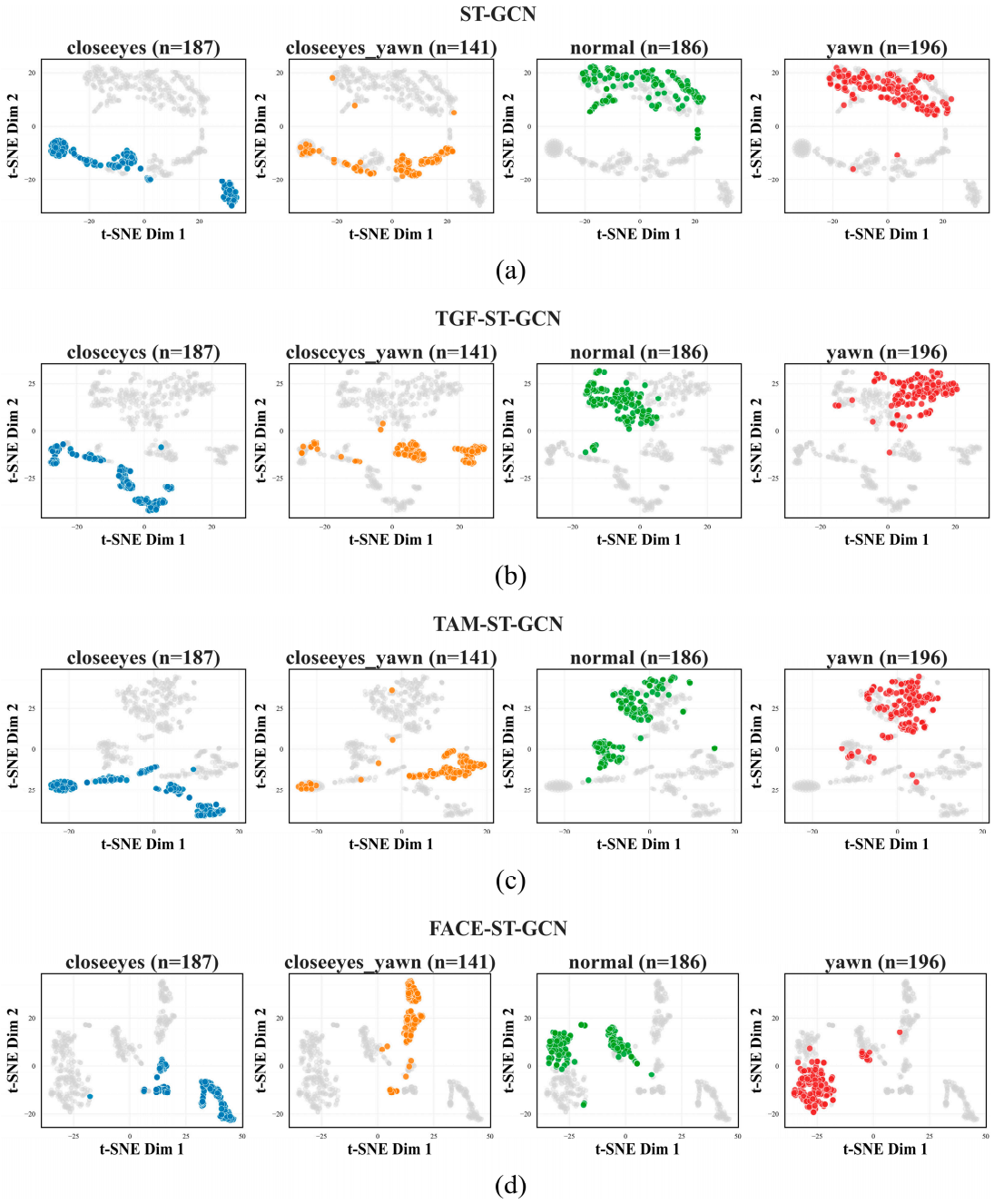


Figure 23. Behavior recognition for different ablation models with t-SNE visualization: (**a**) ST-GCN; (**b**) TGF-ST-GCN; (**c**) TAM-ST-GCN; and (**d**) FACE-ST-GCN.

Misclassified samples are marked with red crosses, and the misclassification results of the four ablation models are illustrated in Figure 24. Figure 24a corresponds to the baseline ST-GCN, which exhibits the largest number of misclassified samples, mainly concentrated in the "yawn," "normal," and "closeeyes" behavior classes, indicating that the original ST-GCN has limited capability in capturing high-dimensional facial behavior features and thus suffers from inferior classification performance. In comparison, TGF-ST-GCN and TAM-ST-GCN, shown in Figure 24b and Figure 24c, respectively, achieve better results, with 53 and 67 misclassified samples, validating the effectiveness of the TGF strategy and the TAM in improving classification performance. As shown in Figure 24d, the proposed FACE-ST-GCN achieves the best performance, further reducing the number of misclassified samples to 37, the lowest among all ablation models. These results demonstrate the complementary advantages of integrating the TGF strategy with the TAM, which jointly enhance the accuracy of facial behavior recognition.

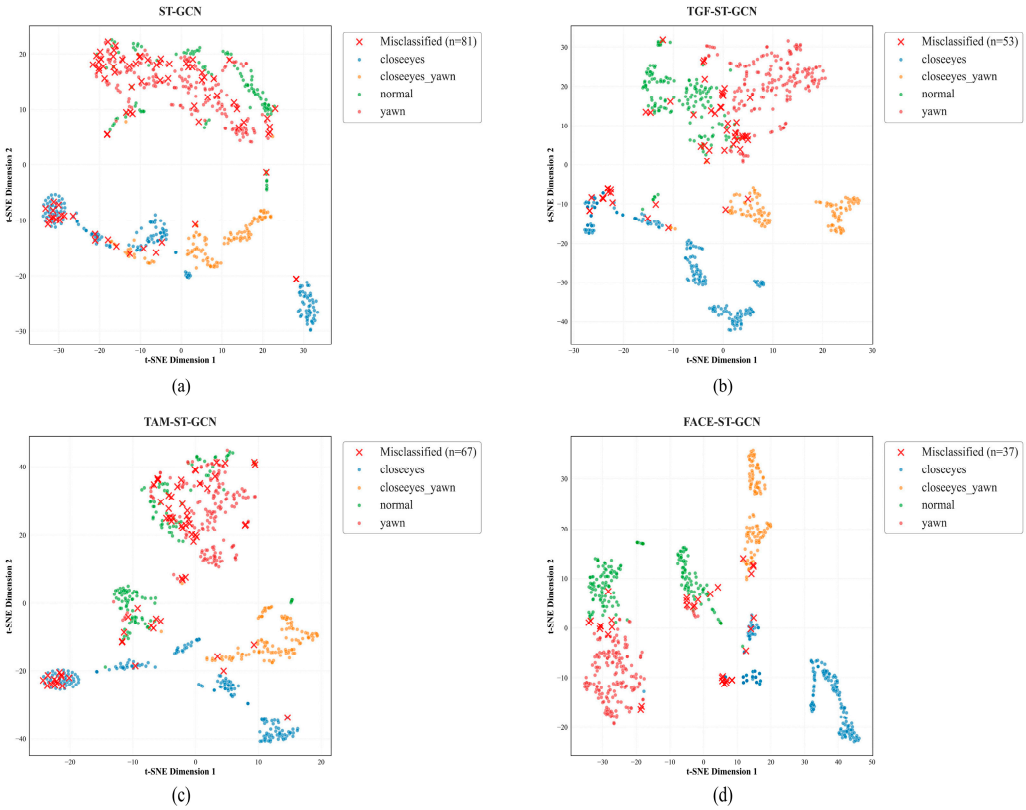


Figure 24. Misclassified samples recognition for different ablation models with t-SNE visualization: (**a**) ST-GCN; (**b**) TGF-ST-GCN; (**c**) TAM-ST-GCN; and (**d**) FACE-ST-GCN.

Table 10 reports the performance comparison of the four ablation models. As shown in the table, although the models incorporating the proposed improvement strategies incur a slight reduction in real-time performance compared with the baseline ST-GCN, they achieve substantial gains in classification accuracy. Specifically, TGF-ST-GCN and TAM-ST-GCN effectively enhance the original model from the spatial and temporal modeling perspectives, respectively, yielding improvements of approximately 10% in *Macro-P*, *Macro-R*, and *Macro-F*1. Building upon these results, FACE-ST-GCN integrates the advantages of both strategies and achieves the best overall performance, with *Macro-P*, *Macro-R*, and *Macro-F*1 improvements of 14.9%, 14.3%, and 14.7% over ST-GCN, respectively. These results convincingly demonstrate the effectiveness of the proposed method in improving classification performance.

Table 10. Ablation experimental results of model performance on the Fatigue-Normal dataset.

Module	*Macro-P*	*Macro-R*	*Macro-F1*	*GFLOPs*	*FPS*
ST-GCN	0.824	0.823	0.822	**5.875**	46.002
TGF-ST-GCN	0.919	0.905	0.906	5.876	**84.718**
TAM-ST-GCN	0.910	0.902	0.904	9.517	50.110
FACE-ST-GCN	**0.947**	**0.941**	**0.943**	9.879	51.254

Note: The bold black font is the best value.

Figure 25 presents the confusion matrices of the compared models. As can be observed, the baseline ST-GCN shows limited capability in distinguishing the four behavior classes, with particularly high misclassification rates for the "normal" and "yawn" behaviors. Specifically, 9.63% and 8.02% of the "normal" samples are misclassified as "closeeyes" and "yawn," respectively, while 1.61% and 24.19% of the "yawn" samples are misclassified as "closeeyes" and "closeeyes_yawn." These results indicate insufficient representation of key facial behavior features by the baseline model. After incorporating the TGF strategy, TGF-ST-GCN achieves a clear improvement in overall recognition accuracy; however, it still exhibits limited discriminative capability for long-duration behaviors such as "yawn." By contrast, TAM-ST-GCN enhances temporal modeling through the TAM, enabling 82.80% of the "yawn" samples to be correctly recognized, although 16.67% are still misclassified as "closeeyes_yawn." Overall, FACE-ST-GCN, which integrates the advantages of both TGF and TAM, delivers the best performance, reducing the misclassification rate of the challenging "yawn" behavior to only 5.38%.

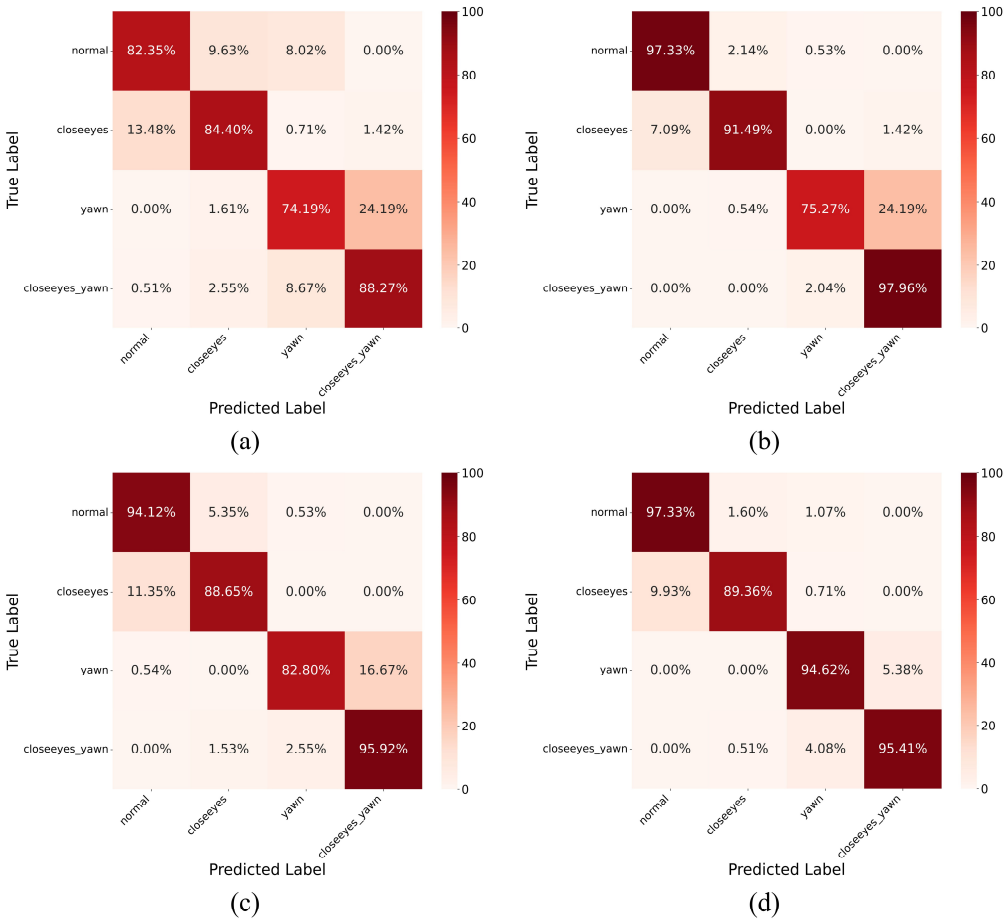


Figure 25. Comparison of confusion matrices for the four ablation models on the Fatigue-Normal dataset: (**a**) ST-GCN; (**b**) TGF-ST-GCN; (**c**) TAM-ST-GCN; and (**d**) FACE-ST-GCN.

Figure 26a–e illustrate the detection results of four ablation models for facial behavior recognition. As shown in the figure, among the four ablation models, the baseline ST-GCN exhibits limited capability in continuous detection of facial behavior variations. Although it achieves a relatively high confidence score for the "closeeyes_yawn" behavior in Figure 26c, its detection confidence across the remaining frames is significantly lower than that of the other models. Moreover, it incorrectly classifies the "normal" state as "closeeyes" in Figure 26d, indicating insufficient temporal modeling ability. In contrast, TGF-ST-GCN, TAM-ST-GCN, and FACE-ST-GCN consistently achieve accurate recognition of all four facial behaviors. Notably, after enhancing temporal modeling through the introduction of the TAM, TAM-ST-GCN demonstrates a substantially higher confidence in detecting the "yawn" behavior compared with TGF-ST-GCN, thereby validating the effectiveness of the TAM in capturing long-term temporal dynamics. FACE-ST-GCN achieves the best detection performance among all compared models. Throughout the process of facial state transitions, it produces no false detections while consistently maintaining high confidence scores, which are the highest among all models. These results further demonstrate that the proposed method effectively integrates the advantages of the TGF strategy and TAM, thereby enabling accurate and robust facial behavior recognition.



Figure 26. Comparison of four ablation models in detecting the face state: (**a**) recognition results for the "closeeyes" state by the model; (**b**) recognition results for the "yawn" state by the model; (**c**) recognition results for the "closeeyes_yawn" state by the model; (**d**) recognition results for the "normal" state by the model; (**e**) recognition results for the " closeeyes " state by the model.

4.6. Experimental Results of the Comparison

To further verify the effectiveness of the proposed model, we conducted comparative experiments against five representative behavior detection frameworks using GCNs: AS-GCN [36], 2S-AGCN [35], Shift-GCN [40], CTR-GCN [41], and Block-GCN [42]. The experiments were conducted on two datasets: 10-Behaviors and Fatigue-Normal. The corresponding results are presented in the following sections.

4.6.1. Comparison Experiments for BODY-ST-GCN

Table 11 reports the evaluation results of different models on the 10-Behaviors dataset. All five models and the proposed BODY-ST-GCN achieve strong detection performance, with *Macro-P*, *Macro-R*, and *Macro-F*1 all exceeding 0.95. Block-GCN attains the best accuracy across these three metrics, but its GFLOPs and *FPS* differ substantially from those of the other methods. By contrast, BODY-ST-GCN yields slightly lower *Macro-P*, *Macro-R* and *Macro-F*1 than Block-GCN, while requiring only 7.411 *GFLOPs*, indicating a lighter computational cost. Correspondingly, BODY-ST-GCN achieves the highest inference speed of 61.503 *FPS* among all compared models. Overall, BODY-ST-GCN maintains high detection accuracy while delivering superior inference efficiency, resulting in a more balanced overall performance.

Table 11. Comparison of experimental results of model performance on the 10-Behaviors dataset.

Module	*Macro-P*	*Macro-R*	*Macro-F*1	*GFLOPs*	*FPS*
AS-GCN	0.963	0.961	0.962	11.751	38.274
2S-AGCN	0.955	0.953	0.954	23.502	19.167
Shift-GCN	0.971	0.970	0.970	11.751	37.555
CTR-GCN	0.954	0.952	0.952	12.371	35.944
Block-GCN	**0.976**	**0.971**	**0.974**	11.752	12.399
BODY-ST-GCN	0.969	0.968	0.969	**7.411**	**61.503**

Note: The bold black font is the best value.

Figure 27 presents the confusion matrices of different models on the 10-Behaviors dataset. As shown, AS-GCN and 2S-GCN exhibit relatively clear main diagonals, indicating that both methods can recognize most categories with good accuracy; however, their performance is weaker on actions with small-amplitude variations, such as “handwaving” and “staggering.” CTR-GCN shows a similar limitation, achieving only 81.36% accuracy on the “handwaving” class. In contrast, Shift-GCN and Block-GCN demonstrate stronger discriminative ability for actions with larger motion amplitudes and more salient pose changes—for example, their accuracies on “fallingdown” and “fighting” both exceed 95%—yet they still underperform on several other categories. Overall, BODY-ST-GCN yields a clear and continuous diagonal pattern, and although its accuracy on a few classes is slightly lower than that of Shift-GCN and Block-GCN, it maintains stable recognition across all behaviors, reflecting good robustness and consistency.

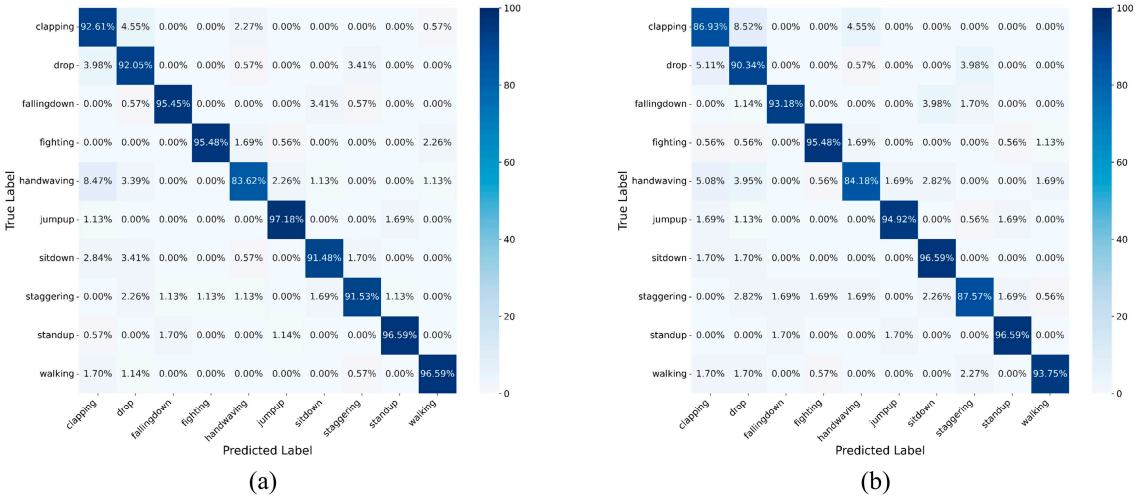


Figure 27. *Cont.*

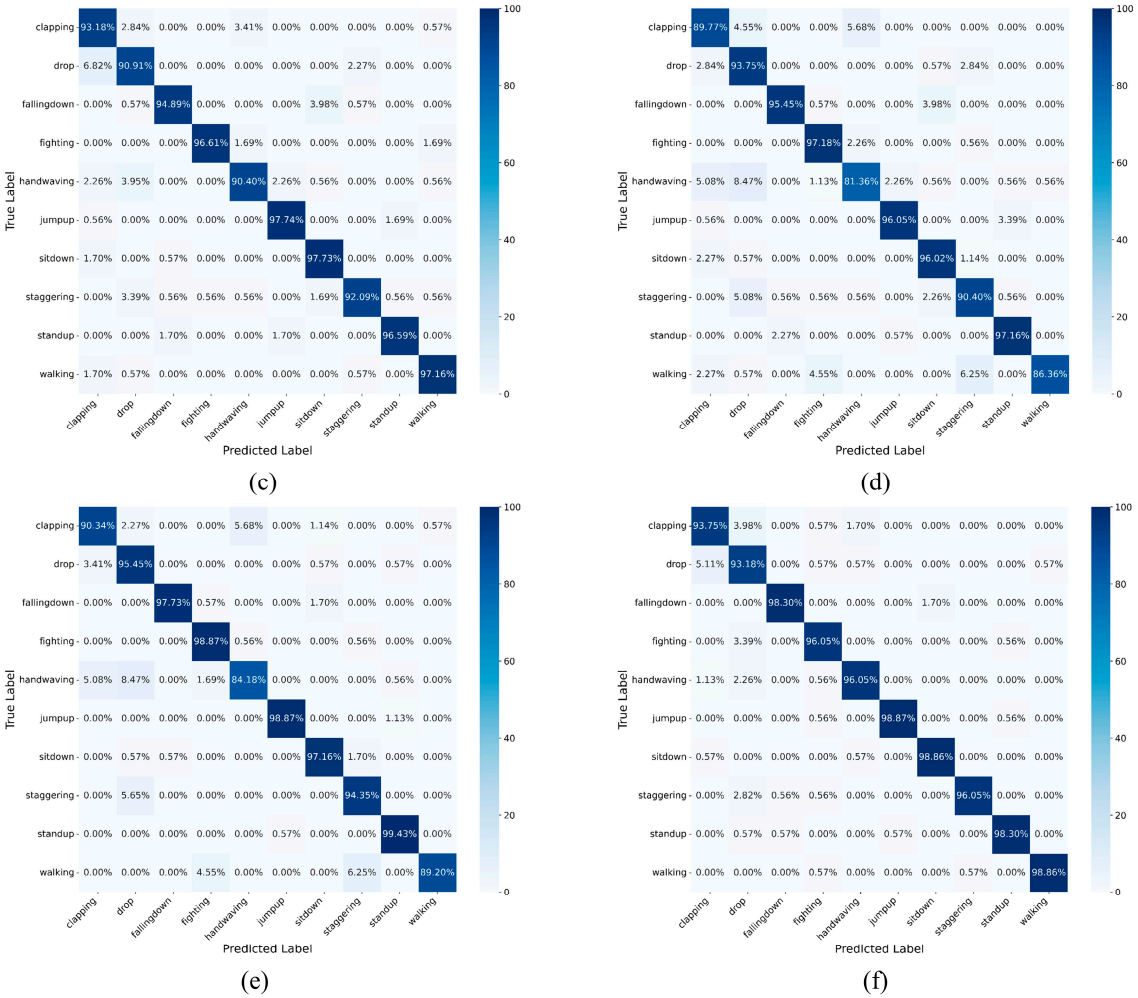


Figure 27. Comparison of confusion matrices for the comparison models on the 10-Behaviors dataset: (**a**) AS-GCN; (**b**) 2S-AGCN; (**c**) Shift-GCN; (**d**) CTR-GCN; (**e**) Block-GCN; and (**f**) Body-ST-GCN.

4.6.2. Comparison Experiments for FACE-ST-GCN

Table 12 presents the performance evaluation of different models on the Fatigue-Normal dataset. As shown in the table, Block-GCN achieves the best detection performance among all compared methods, obtaining the highest values in terms of *Macro-P*, *Macro-R*, and *Macro-F*1. However, Block-GCN exhibits relatively high *GFLOPs* and low *FPS*, indicating limited real-time capability for this task. In contrast, FACE-ST-GCN achieves slightly lower *Macro-P*, *Macro-R*, and *Macro-F*1 scores than Block-GCN, while significantly outperforming it in terms of *GFLOPs* and *FPS*, ranking first among all models in computational efficiency. Overall, FACE-ST-GCN strikes a favorable balance between detection accuracy and real-time performance.

Table 12. Comparison of experimental results of model accuracy on the Fatigue-Normal dataset.

Module	*Macro-P*	*Macro-R*	*Macro-F*1	*GFLOPs*	*FPS*
AS-GCN	0.746	0.719	0.718	18.571	29.992
2S-AGCN	0.844	0.835	0.837	43.398	18.575
Shift-GCN	0.877	0.863	0.866	45.699	20.799
CTR-GCN	0.819	0.804	0.804	38.524	18.525
Block-GCN	**0.950**	**0.944**	**0.946**	15.701	27.660
FACE-ST-GCN	0.942	0.936	0.938	**9.879**	**51.254**

Note: The bold black font is the best value.

Figure 28 illustrates the confusion matrices of the evaluated models. As shown in the figure, AS-GCN exhibits the weakest performance in detecting the "yawn" behavior, with only 45.16% of "yawn" samples correctly classified, while 52.69% are misclassified as "closeeyes_yawn". In addition, AS-GCN performs poorly in distinguishing between the "closeeyes" and "closeeyes_yawn" behaviors. In comparison, 2S-AGCN, Shift-GCN, and CTR-GCN achieve improved recognition accuracy for actions such as "yawn", but still suffer from misclassification rates of approximately 10–20%. By contrast, Block-GCN and FACE-ST-GCN substantially alleviate these issues, significantly improving the detection accuracy of both "yawn" and "closeeyes_yawn", and increasing the recognition accuracy of the "closeeyes" behavior to over 85%.

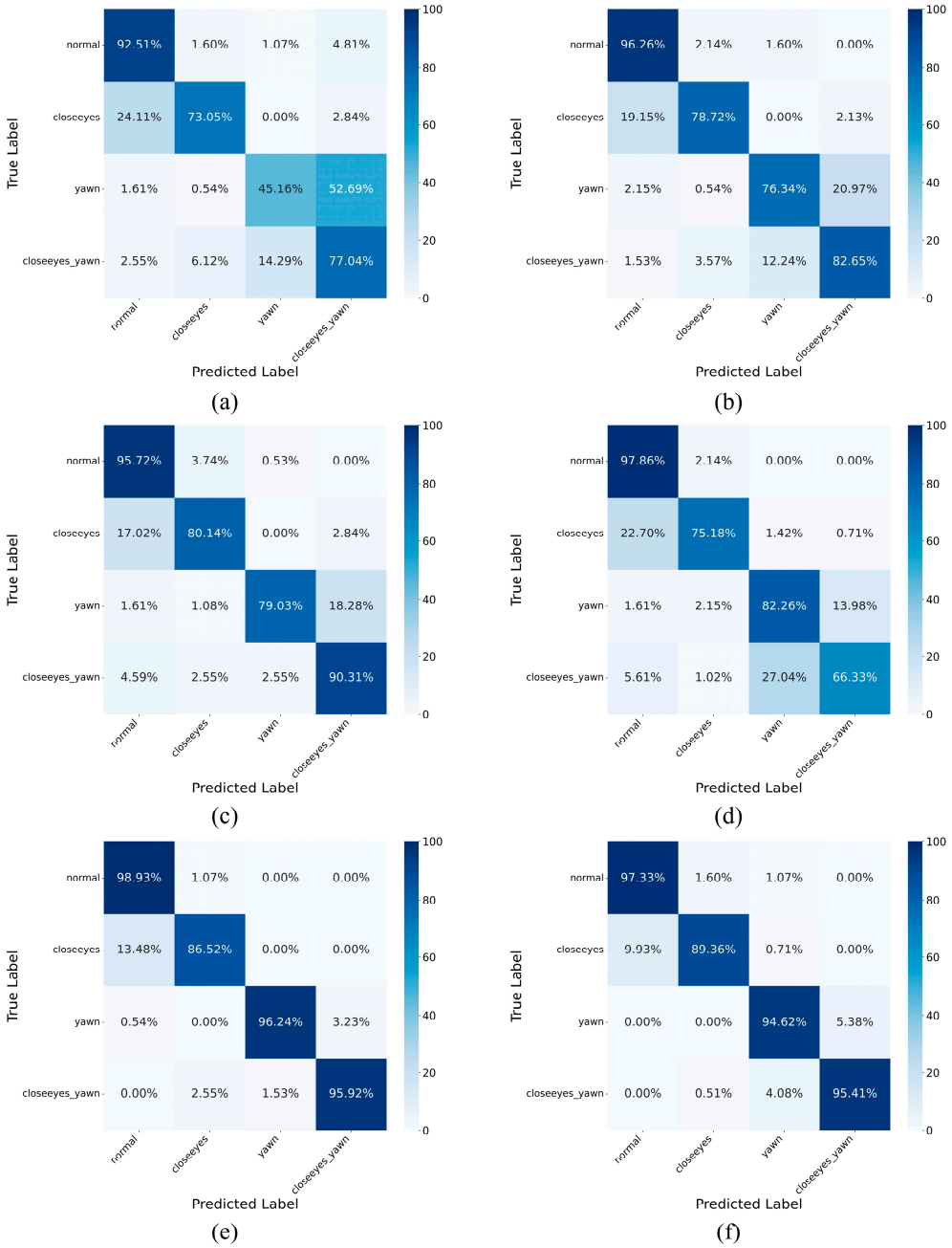


Figure 28. Comparison of confusion matrices for the comparison models on the Fatigue-Normal dataset: (**a**) AS-GCN; (**b**) 2S-AGCN; (**c**) Shift-GCN; (**d**) CTR-GCN; (**e**) Block-GCN; and (**f**) FACE-ST-GCN.

4.7. Visualized Results of the Joint Detection System

Figure 29 presents the behavior state detection results over a continuous temporal sequence. The upper part shows the original video frames, while the lower part illustrates the corresponding detection and recognition results. As observed, the human posture remains in a seated position throughout the entire sequence, whereas the facial state varies dynamically over time. Specifically, during the interval from T = 10 s to T = 60 s, the eye region remains continuously closed, and yawning behavior begins to appear from T = 40 s. According to the predefined risk state criteria, this period is classified as state S1. Based on the eye fatigue detection standard established in this paper, the system identifies prolonged eye closure at T = 60 s, indicating that the subject may have entered a fatigued state, and the corresponding state is therefore updated to S2. Subsequently, during the interval from T = 70 s to T = 80 s, the facial state gradually returns to normal. However, since the subject maintains a seated posture, the system continues to classify the state as S2.



Figure 29. Joint detection results for OOW behaviors and facial states.

Overall, the proposed dual-branch ST-GCN model achieves accurate and stable detection of both human posture and facial state over continuous temporal sequences. During the transitional phases of facial states, no evident false positives or missed detections are observed. Furthermore, under the joint risk assessment criteria, the system is able to

provide accurate classification of the current risk state while maintaining a high level of confidence in behavior recognition.

5. Conclusions

To address the limitations of traditional ST-GCN in modeling complex behaviors and its inability to effectively balance body movements and facial fatigue features, this paper proposes a dual-branch ST-GCN system for detecting OOW unsafe behaviors and facial fatigue states. The algorithm first uses OpenPose to extract skeleton and facial keypoints for OOW detection. These keypoints are then fed into the BODY-ST-GCN and FACE-ST-GCN branches to perform recognition of OOW behaviors and facial features, respectively. Specifically, the BODY-ST-GCN enhances spatial modeling capability using a TGF strategy and incorporates TSP-TCN to capture multi-scale temporal features. In the FACE-ST-GCN, the GCN part also uses the TGF strategy, while the TCN part introduces the TAM to enable the model to focus on key temporal information related to fatigue-related facial features, thereby improving the recognition accuracy of facial features. In addition, a hierarchical risk assessment mechanism is designed by jointly considering body behaviors and facial fatigue features, enabling the classification of OOW duty states into multiple risk levels. This mechanism supports continuous and dynamic evaluation of OOW status, further improving the interpretability and practical value of the proposed system.

To validate the effectiveness of the algorithm, ablation experiments are conducted on the 10-Behaviors dataset derived from the NTU-RGB+D-60 dataset, and the Fatigue-Normal dataset from the YawDD dataset. Ablation experiments on the 10-Behaviors dataset and the Fatigue-Normal dataset show that on the 10-Behaviors dataset, BODY-ST-GCN achieves *Macro-P*, *Macro-R*, and *Macro-F*1 scores of 0.969, 0.968, and 0.969, respectively, improving by 6.4%, 6.6%, and 6.8% compared to the baseline model, ST-GCN. On the Fatigue-Normal dataset, FACE-ST-GCN reaches *Macro-P*, *Macro-R*, and *Macro-F*1 scores of 0.942, 0.936, and 0.938, respectively, with improvements of 14.5%, 14.0%, and 14.3% over the baseline model. Additionally, compared to five advanced detection models, the two proposed ST-GCN demonstrated superior performance in terms of overall accuracy and the ability to detect various behaviors. These results demonstrate the effectiveness of the proposed dual-branch ST-GCN system in detecting OOW unsafe behaviors and facial fatigue features during duty shifts.

However, the shipboard validation dataset constructed in this study still has certain limitations. Since the experimental data are primarily collected under controlled conditions through simulated scenarios, and the number of participants is relatively limited, the dataset may not fully capture the complexity and randomness of real shipboard environments. In addition, the proposed model still suffers from relatively high computational complexity. To address these issues, future work will focus on the following directions: (1) designing lightweight network architectures to improve real-time performance and computational efficiency; (2) expanding the scale and diversity of the shipboard dataset by collecting more natural behavioral data from real duty scenarios and further enhancing the model's adaptability to complex environmental conditions by incorporating factors such as facial occlusion, lateral pose variations, and complex or low-light illumination commonly encountered on the ship bridge; and (3) enhancing the model's capability under complex environmental conditions to improve its stability and generalization in practical applications. Ultimately, this will enable the development of an intelligent monitoring system that operates effectively across all weather conditions and diverse scenarios.

Author Contributions: Conceptualization, R.Q. and S.X.; methodology, R.Q. and S.X.; software, R.Q. and S.X.; validation, R.Q., S.X., and K.C.; formal analysis, R.Q. and S.X.; investigation, R.Q. and S.X.; resources, R.Q. and S.X.; data curation, R.Q., S.X., K.C., Z.Z., and X.H.; writing—original draft

preparation, R.Q.; writing—review and editing, R.Q., S.X., and K.C.; visualization, R.Q. and S.X.; supervision, S.X., K.C., Z.Z., and X.H. project administration, R.Q. and S.X.; funding acquisition, S.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [Grant No. 52231014], the Liaoning Provincial Shipping Joint Fund [Grant No. 2020HYLH-27], and the Fundamental Research Funds for the Central Universities [Grant No. 3132024618].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data will be made available on request.

Acknowledgments: We would like to thank Zijian Zhang and Xiaoyu He for their support in data curation and project supervision.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. European Maritime Safety Agency. *Annual Overview of Marine Casualties and Incidents 2023*; European Maritime Safety Agency: Lisbon, Portugal, 2023. Available online: https://www.emsa.europa.eu/publications/reports/item/5052-annual-overview-of-marine-casualties-and-incidents.html (accessed on 12 February 2026).
2. Shi, X.; Zhuang, H.; Xu, D. Structured Survey of Human Factor-Related Maritime Accident Research. *Ocean Eng.* **2021**, *237*, 109561. [CrossRef]
3. *MSC.128(75)*; Performance Standards for a Bridge Navigational Watch Alarm System (BNWAS). International Maritime Organization: London, UK, 2002. Available online: https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/MSCResolutions/MSC.128(75).pdf (accessed on 12 February 2026).
4. Cheng, K.W.E.; Xue, X.D.; Chan, K.H. Zero emission electric vessel development. In Proceedings of the 2015 6th International Conference on Power Electronics Systems and Applications (PESA), Hong Kong, China, 15–17 December 2015. [CrossRef]
5. Gomes, A.; Ke, W.; Lm, S.K.; Siu, A.; Mendes, A.J.; Marcelino, M.J. A teacher's view about introductory programming teaching and learning—Portuguese and Macanese perspectives. In Proceedings of the 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA, 18–21 October 2017. [CrossRef]
6. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1958–1974. [CrossRef] [PubMed]
7. Liu, Z.; Lyu, K.; Wu, S.; Chen, H.; Hao, Y.; Ji, S. Aggregated Multi-GANs for Controlled 3D Human Motion Prediction. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2225–2232. [CrossRef]
8. Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; Tran, D. Detect-and-Track: Efficient Pose Estimation in Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 350–359. [CrossRef]
9. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018. [CrossRef]
10. International Maritime Organization. *Amendments to the Code for the Investigation of Marine Casualties and Incidents, A.884(21)*; International Maritime Organization: London, UK, 1999. Available online: https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.884(21).pdf (accessed on 9 February 2026).
11. Sant'Ana, M.; Guo, L.; Hou, Z. A Decentralized Sensor Fusion Approach to Human Fatigue Monitoring in Maritime Operations. In Proceedings of the 2019 IEEE 15th International Conference on Control and Automation (ICCA), Edinburgh, UK, 16–19 July 2019; pp. 1569–1574. [CrossRef]
12. Monteiro, T.G.; Skourup, C.; Zhang, H. A Task Agnostic Mental Fatigue Assessment Approach Based on EEG Frequency Bands for Demanding Maritime Operations. *IEEE Instrum. Meas. Mag.* **2021**, *24*, 82–88. [CrossRef]
13. Li, C.; Fu, Y.; Ouyang, R.; Liu, Y.; Hou, X. ADTIDO: Detecting the Tired Deck Officer with Fusion Feature Methods. *Sensors* **2022**, *22*, 6506. [CrossRef] [PubMed]
14. Youn, I.-H.; Park, D.-J.; Yim, J.-B. Analysis of Lookout Activity in a Simulated Environment to Investigate Maritime Accidents Caused by Human Error. *Appl. Sci.* **2018**, *9*, 4. [CrossRef]
15. Chen, M.; Zhang, L.; Liu, Y.; Zhang, Y.; Liu, C.; Chen, M. Ship Bridge OOW Activity Status Detection Using Wi-Fi Beamforming Feedback Information. *J. Mar. Sci. Eng.* **2024**, *12*, 872. [CrossRef]
16. Lin, X.; Wang, S.; Sun, Z.; Zhang, M. YOLO-SD: A Real-Time Crew Safety Detection and Early Warning Approach. *J. Adv. Transp.* **2021**, *2021*, 7534739. [CrossRef]

17. Zhao, C.; Zhang, W.; Chen, C.; Yang, X.; Yue, J.; Han, B. Recognition of Unsafe Onboard Mooring and Unmooring Operation Behavior Based on Improved YOLO-v4 Algorithm. *J. Mar. Sci. Eng.* **2023**, *11*, 291. [CrossRef]
18. Liu, W.; Liu, X.; Hu, Y.; Shi, J.; Chen, X.; Zhao, J.; Wang, S.; Hu, Q. Fall Detection for Shipboard Seafarers Based on Optimized BlazePose and LSTM. *Sensors* **2022**, *22*, 5449. [CrossRef] [PubMed]
19. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-Person Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112. [CrossRef]
20. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 472–487. [CrossRef]
21. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite Fields for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11969–11978. [CrossRef]
22. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5385–5394. [CrossRef]
23. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef] [PubMed]
24. Wang, C.; Yan, J. A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. *IEEE Access* **2023**, *11*, 53880–53898. [CrossRef]
25. Du, Y.; Wang, W.; Wang, L. Hierarchical Recurrent Neural Network for Skeleton-Based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1110–1118. [CrossRef]
26. Cui, R.; Zhu, A.; Zhang, S.; Hua, G. Multi-Source Learning for Skeleton-Based Action Recognition Using Deep LSTM Networks. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 547–552. [CrossRef]
27. Liu, J.; Wang, G.; Duan, L.-Y.; Abdiyeva, K.; Kot, A.C. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. *IEEE Trans. Image Process.* **2018**, *27*, 1586–1599. [CrossRef] [PubMed]
28. Liu, H.; Tu, J. Two-Stream 3D Convolutional Neural Network for Human Skeleton-Based Action Recognition. *arXiv* **2017**, arXiv:1705.08106. [CrossRef]
29. Wang, P.; Li, W.; Li, C.; Hou, Y. Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks. *Knowl. Based Syst.* **2018**, *158*, 43–53. [CrossRef]
30. Ding, Z.; Wang, P.; Ogunbona, P.O.; Li, W. Investigation of Different Skeleton Features for CNN-Based 3D Action Recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 617–622. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
32. Qiu, H.; Hou, B.; Ren, B.; Zhang, X. Spatio-Temporal Tuples Transformer for Skeleton-Based Action Recognition. *arXiv* **2022**, arXiv:2201.02849. [CrossRef]
33. Wu, W.; Wang, P.; Chen, C.; Lu, A. FreqMixFormerV2: Lightweight Frequency-aware Mixed Transformer for Human Skeleton Action Recognition. In Proceedings of the 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG), Clearwater, FL, USA, 12–16 May 2025; pp. 1–5. [CrossRef]
34. Wu, W.; Guo, Z.; Chen, C.; Lu, A. UniSTFormer: Unified Spatio-Temporal Lightweight Transformer for Efficient Skeleton-Based Action Recognition. *arXiv* **2025**. [CrossRef]
35. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12026–12035. [CrossRef]
36. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-Based Behavior Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
37. Liu, Z.; Wang, L.; Wu, W.; Qian, C.; Lu, T. TAM: Temporal Adaptive Module for Video Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13688–13698. [CrossRef]
38. Shahroudy, A.; Liu, J.; Ng, T.-T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019. [CrossRef]

39. Omidyeganeh, M.; Shirmohammadi, S.; Abtahi, S.; Khurshid, A.; Farhan, M.; Scharcanski, J.; Hariri, B.; Laroche, D.; Martel, L. Yawning Detection Using Embedded Smart Cameras. *IEEE Trans. Instrum. Meas.* **2016**, *65*, 570–582. [CrossRef]
40. Cheng, K.; Zhang, Y.; He, X. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 180–189. [CrossRef]
41. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13339–13348. [CrossRef]
42. Zhou, Y.; Yan, X.; Cheng, Z.Q.; Yan, Y.; Dai, Q.; Hua, X.S. BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 2049–2058. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.