IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 June 2024, accepted 5 August 2024, date of publication 12 August 2024, date of current version 22 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3442371

RESEARCH ARTICLE

A Class Balanced Spatio-Temporal Self-Attention Model for Combat Intention Recognition

XUAN WANG[1], BENZHOU JIN[1], (Member, IEEE), MINGYANG JIA[1], GANG WU[2], AND XIAOFEI ZHANG[1]

[1]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[2]14th Research Institute, China Electronics Technology Group Corporation, Nanjing 211106, China

Corresponding author: Benzhou Jin (jinbz@nuaa.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62371230, and in part by Jiangsu Provincial Key Research and Development Program under Grant BE2023027.

ABSTRACT To address the issue of model performance degradation in combat intention recognition caused by the long-tailed distribution of battlefield data and the neglect of the spatial dimension information of multivariate time series data, this paper proposes a class balanced spatio-temporal self-attention (CBSTSA) model. By incorporating spatial and temporal attention mechanisms, the model captures interdependencies among features and extracts salient information from both temporal and spatial dimensions. Furthermore, taking the long-tailed distribution of battlefield data into account, a re-weighted class balanced loss function is introduced to train the model. Experimental results show the superiority of our CBSTSA model, e.g. achieving approximately 95.67% accuracy in typical scenarios, surpassing benchmark schemes by 4–5%.

INDEX TERMS Combat intention recognition, long-tailed distribution, self-attention.

I. INTRODUCTION

Combat intention recognition refers to the judgment and interpretation of the enemy's operational assumptions, and operational plans by analyzing the information obtained from various information sources on the battlefield. According to the different level of impact of combat, it can be subdivided into strategic intention recognition, campaign intention recognition and tactical intention recognition [1]. This paper focuses on tactical intention recognition.

The intent recognition methods can be generally divided into two categories: model-driven methods and data-driven methods [2]. With the development of science and technology in recent years, the forms of modern warfare and elements of confrontation have become more and more complex, and the amount of data that needs to be processed in the battlefield has increased exponentially. Model-driven methods are difficult to meet the needs of today's battlefield, thus more and more researchers focused on data-driven intention recognition methods in recent years, e.g., neural network and deep learning [3], [4], [5], [6], [7], [8].

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

The majority of the current data-driven methods treat the combat intention recognition task as a multi-classification problem, which based on multivariate time series data. These approaches commonly employ recurrent neural networks (RNNs) [9] and their variant [10], or temporal convolutional networks (TCNs) [5] to primarily extract temporal features. In [3], [4], [7], and [8], the authors combined convolutional neural networks (CNNs) [11] and RNNs in joint architectures to capture both local and global dependencies in the temporal dimension. However, in a real combat scenario, the feature obtained by sensors should be viewed as a two-dimensional space that encompasses both spatial (the correlation between different features attributes at a single time) and temporal dimensions. The above models only analyze data from temporal dimension, disregarding the valuable information embedded in the spatial dimension. As a consequence, a significant number of essential features within the intention samples are overlooked.

Moreover, a significant limitation of the current data-driven methods is the lack of consideration for the imbalanced distribution of battlefield situation data across different intentions. Actually, intentions with low threat levels, such as scout and early warning, can collect a large number of

samples in the real battlefield. In contrast, some high-threat intentions, such as strikes and intercepts, tend to appear less frequently and with a shorter duration in the battlefield, resulting in a smaller sample size. This characteristic lead to a special skewed distribution within the dataset, known as a long-tailed distribution [12]: a few categories, termed head classes, dominate the number of samples, while others, referred to as tail classes, have only a small number of samples. The long-tailed distribution can lead to the data-driven model performing well in the head classes but inefficient for tail classes, thereby negatively impacting the overall recognition accuracy. The tail classes in combat intention recognition usually correspond to the intentions with high threat levels, which play a more critical role in supporting the commander's operational command. Failure to enhance performance on these tail classes could significantly undermine the reliability of the entire intention recognition system could.

In this paper, we propose a novel class balanced spatio-temporal self-attention (CBSTSA) model for the combat intention recognition task. Extensive solid results show that CBSTSA outperforms benchmark schemes, including GRU-FCN [3], 1DCNN-BiLSTM [4], TCN-attention [5], etc. The main contributions are as follows:

(1) A spatio-temporal self-attention (STSA) mechanism is proposed to enhance feature extraction. The STSA mechanism begins by transposing the time series data, enabling the subsequent application of the spatial multi-head self-attention module. The spatial module serves to enhance the features of the original time series data in the spatial dimension. Additionally, a time multi-head self-attention module is employed to capture long-term features in the time dimension. The proposed space-time tandem structure facilitates the holistic utilization of the original data in both temporal and spatial domains, while maintaining the inherent time dimension structure of the input data.

(2) A novel class balanced (CB) loss function is proposed. Aiming at the performance degradation of the intention recognition models caused by long-tailed distributions, we employ the novel CB loss function during training. The proposed CB loss function integrates two key factors: the sample size and the threat level. This function is designed with a reweighting term that balances the trade-off between these factors for each class. By assigning a strategic penalty, the model is encouraged to focus on classes with fewer samples and higher threat levels. This approach helps the model to better capture the distinctive features of these classes, thereby enhancing its performance on imbalanced class distributions.

The rest of this paper is organized as follows. Section II provides a summary of the related work.[1] Section IV presents the problem statement and the details of the CBSTSA model. Section V includes experimental validation and result analysis. Section VI concludes the work of this paper.

[1]Section I discusses the overall situation, existing problems, and development trends of the research in combat intention recognition field, and indicates contributions of our works on this basis. Details and analysis of related works will be presented in Section II.

II. RELATED WORKS

Most of previous efforts on combat intention recognition can be divided into two regimes: model-driven and data-driven [2].

A. MODEL-DRIVEN

Model-driven methods typically involve the predefined models based on prior knowledge or expert experience. These methods aim to explicitly model the mapping relationship between the input and output through various techniques, including template matching [13], [14], [15], [16], expert systems [17], [18], [19], Bayesian networks [20], [21], and D-S evidence theory [22].

The intention recognition method based on template matching was initially proposed by Azarewicz et al. [13], which constructed an intention recognition model by integrating predictions of future activities and the assumptions of external behavior recognition. However, with the increasing volume of data on the battlefield, it becomes increasingly challenging to construct templates. The expert system, an important branch of early artificial intelligence, was first applied to combat intention recognition by Kirillov et al. [17]. They transformed the intention recognition problem into a multi-hypothetical dynamic classification problem and achieved identification of the threat target intention by continuously received data. This method is more flexible than template matching, but it heavily relies on prior knowledge and subjective factors, resulting in poor robustness of the models. Subsequent studies have made improvements to template matching [14], [15], [16] and expert systems [18], [19] for various problems. Despite these improvements enhancing the performance and fault tolerance of the model, it still retains the inherent limitations of the methods. These limitations make template matching and expert system challenging to handle uncertain reasoning and adapt to the increasingly complex battlefield environment.

Due to the inherent capability of Bayesian networks and D-S evidence theory in dealing with uncertain problem, they have been employed for combat intention recognition task. Bayesian networks, for instance, combine prior probabilities and uncertainty reasoning to achieve improved analysis results. They can address the challenges of uncertainty and incompleteness by deriving the output probabilities based on input variable information. Deng et al. [20] proposed the use of a multi-entity Bayesian network (MEBN) to describe the tactical intention of the enemy. The MEBN extends Bayesian networks using first-order predicate logic to accommodate multi-entity combat situations, but it neglects the temporal dimension of information. Yu [21] analyzed the enemy's intentions from the time dimension, introducing a dynamic Bayesian network (DBN) and incorporating fuzzy classification functions and probability conversion theory to reduce errors arising from subjective judgments. However,

the difficulty of constructing the DBN will continuously increase with the growing number of battlefield factors. Compared with Bayesian networks, D-S evidence theory is capable of handling weaker conditions and excels in expressing and synthesizing uncertain information. Sun et al. [22] combined the high-dimensional spatial similarity calculation model with the D-S evidence theory to achieve sequential identification of the target's tactical intention. Although Bayesian network and D-S evidence theory address the limitations of previous methods in dealing with uncertainty, as probabilistic inference models, they still struggle to avoid the influence of subjective assumptions on the model, such as setting judgment thresholds and prior probabilities. Furthermore, with the continuous growth in the volume of battlefield data, manually constructing these models becomes increasingly challenging. As a result, the aforementioned methods cannot fully meet the requirements of current combat intention recognition tasks.

B. DATA-DRIVEN

Data-driven methods involve the automatic construction of inference models by learning the mapping between input and output data. In the context of combat intention recognition, most of data-driven models treat it as a time series data classification task. Ou et al. [6] designed a tactical intent recognition model based on stacked auto encoder (SAE), which encodes data from multiple moments uniformly as network inputs to achieve intention recognition. The SAE model is not time-sensitive, therefore the approach of encoding-recognizing is not intuitive. Liu et al. [7] proposed a CNN-GRU model for intention recognition by considering the combination of static map features and action trajectories of combat units in a war game. The model consists of a 2-dimensional CNN for extracting map features and a 1-dimensional CNN for extracting temporal features, but it is specific to the war game scenario and difficult to apply in practical settings. Li et al. [8] proposed a hierarchical aggregation model based on CNN-BiLSTM-attention, taking into account the hierarchical nature of combat intention and the dependency relationship between intent behaviors. This method is designed for real combat environments but still relies on ideal assumptions, overlooking the complex situations present on the battlefield, such as inconsistent costs of misjudgment and the imbalanced sample distributions. To address the problem of imbalanced battlefield intent misjudgment costs, Ding et al. [3] proposed a cost-sensitive aerial target intention recognition method based on GRU-FCN. Zhang et al. [4] divided the target data into multiple sub-sequences and employed a baseline model called 1DCNN-BiLSTM to handle long time series data. Taking the sample imbalance into consideration, Zhao et al. [5] proposed a temporal convolutional self-attention network (TCN-attention) based on sliding-window estimation. The authors addressed the imbalance by determining difficult-to-categorize samples through pre-training and oversampling those categories through sliding-window estimation expansion, thereby alleviating the sample number imbalance. However, all the aforementioned methods neglect the spatial dimension of the battlefield, which limits the improvement in performance.

III. PROBLEM FORMULATION

Data-driven approaches usually model the combat intention recognition task as a multivariate time series classification problem. A single moment of intent feature $\boldsymbol{x}$ is first determined based on the combat scenario, which can be formally represented as:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \tag{1}$$

where x_n denotes the n-th intentional characteristics of the enemy target at a given moment, which is then extended to m consecutive moments to obtain a complete representation of feature space $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^{t_1} \\ \boldsymbol{x}^{t_2} \\ \vdots \\ \boldsymbol{x}^{t_m} \end{bmatrix} = \begin{bmatrix} x_1^{t_1} & x_2^{t_1} & \cdots & x_n^{t_1} \\ x_1^{t_2} & x_2^{t_2} & \cdots & x_n^{t_2} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{t_m} & x_2^{t_m} & \cdots & x_n^{t_m} \end{bmatrix} \tag{2}$$

where $x_n^{t_m}$ denotes the n-th intentional characteristics of the enemy target at the moment t_m.

The enemy targets' intent space is determined as:

$$I = \{i_1, i_2, \ldots, i_p\} \tag{3}$$

where i_p denotes the p-th intention of the target.

The mapping relation $f(\cdot)$ from the feature space $\boldsymbol{X}$ to the intent space I, which the data-driven models need to learn, can be expressed as:

$$I = f(\boldsymbol{X}) = f\left(\left[\boldsymbol{x}^{t_1}, \boldsymbol{x}^{t_2}, \ldots, \boldsymbol{x}^{t_m}\right]^T\right) \tag{4}$$

In this paper, we focus on a typical combat scenario where our battle units confront enemy targets. Our goal is to recognize the intention of the enemy target based on the information obtained by our sensors and the current state of our battle units. Considering the specific task, it is observed that the same intention often exhibits similar characteristics. By considering these characteristics patterns, the state information of our battle units, the radiation source category information equipped by the enemy target and the radiation source working mode information are determined as the features of a single moment.

Based on the above discussion, we construct a feature space with a time dimension of m simulation moments for the above sea combat scenario. Each moment in the feature space consists of five feature parameters: distance between ourselves and the enemy (d), speed of our battle units (v), altitude of our battle units (h), the type of enemy target's radiation source ($\boldsymbol{t}$), and the working mode of the radiation source ($\boldsymbol{w}$). Table 1 presents the specific form of these feature parameters. The feature space consists of two main types of data, numerical and non-numerical. The numerical

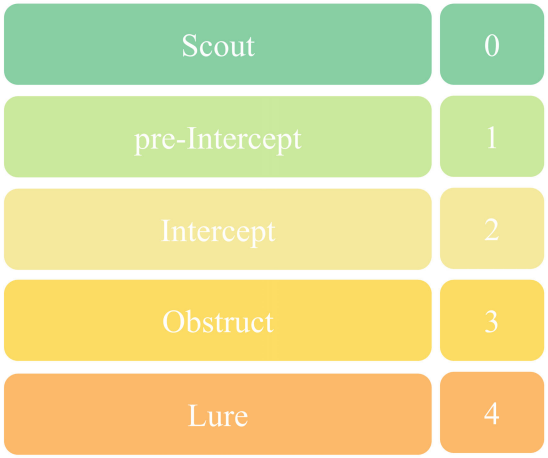

FIGURE 1. Intention space.

parameters include: distance between ourselves and the enemy (d), speed of our battle units (v), altitude of our battle units (h), and the non-numerical parameters includes: the type of enemy target's radiation source ($\boldsymbol{t}$), and the working mode of the radiation source ($\boldsymbol{w}$). The non-numerical parameters are represented using multi-hot encoding with a total of 12-bit coding. Therefore, the feature space $\boldsymbol{X} \in \mathbb{R}^{m\times 15}$, where m represents the number of simulation moments. By constructing this feature space, we capture essential information about the sea combat scenario, enabling the subsequent intention recognition process.

The intent space is generally determined by experts in warfare. In this paper, we establish an output space that encompasses five types of intentions: scout, pre-intercept, intercept, obstruct, and lure. The determination of these intentions is based on a comprehensive consideration of factors such as the battlefield environment of the multiple battle units cooperative combat, the working state of the enemy's radiation sources, and the mission context. Fig. 1 represents the encoding and representation of these five types of intentions within the intent space.

IV. APPROACH

A. OVERVIEW

In order to extract the spatial features of the battlefield and to address the long-tailed distribution issue, we propose CBSTSA model for combat intention recognition. An overview of our CBSTSA model can be found in Fig. 2. The CBSTSA model consists of spatio-temporal self-attention (STSA) mechanism and class balanced (CB) loss function. More specifically, we start by training our model on the training set using the CB loss, as indicated by the yellow arrow in the diagram. This approach allows our model to focus more on fitting the few-shot, high-threat categories effectively. And then, the trained model can be used for testing. During test, the intention is directly obtained by applying argmax function to the output of STSA, as indicated by the orange arrow in the diagram.

B. SPATIO-TEMPORAL SELF-ATTENTION MECHANISM

Multivariate time series refers to a time series dataset containing multiple observed variables at a single moment [23]. When analyzing multivariate time series, we should not only focus on the temporal dependence of these variables (temporal dimension), but also pay attention the connection of observed variables with each other (spatial dimension). Currently, the dominant combat intention recognition models are based on the structure of 1DCNNs+RNNs (in series [4], [7], [8] or in parallel [3]). The 1DCNN module performs convolutional operations on the temporal dimension of the data. This allows it to extract more refined local features in the time domain. And the RNN module, with its ability to model contextual dependencies, efficiently captures the long-term dependency features of the time series data.

However, the above structure does not consider the connection among observations of different features, i.e., the features of spatial dimensions. To address this challenge, this paper proposes a novel spatio-temporal self-attention (STSA) mechanism for intention recognition. The proposed model comprises three main components: a spatial attention module with n_s self-attention mechanisms, a temporal attention module with n_t self-attention mechanisms, and a fully connected layer with *softmax* activation function (serves as the classification layer). The architecture of the model is depicted in Fig. 3.

Instead of analyzing the battlefield situation data only in the temporal dimension, the proposed spatio-temporal self-attention (STSA) augments the input data in the spatial dimension before extracting features from the temporal dimension. Specifically, the spatial attention module takes the transposed time series data $\boldsymbol{X}^T$ as input, and aims to capture the dependencies among the variables from the spatial domain. Through this process, the module obtains the enhanced time series data feature $\boldsymbol{f}_{space}$, which encompasses refined information from the spatial dimension. Subsequently, the temporal attention module receives the transposed output of spatial attention module, denoted as $\boldsymbol{f}^T_{space}$, as its input. The temporal attention module, similar to RNNs, focuses on capturing the long-term dependencies within the data from the temporal dimension. This pipeline enables the model to extract more comprehensive and meaningful temporal features which play a critical role in enhancing the understanding and recognition of intentions. Finally, the final classification result is obtained by the classification layer, which is a fully connected layer with softmax activation.

The basic unit of both the spatial attention module and the temporal attention module is the multi-head self-attention mechanism, as shown in Fig. 4. Each self-attention module can be represented as:

$$Attention\,(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \alpha\boldsymbol{V} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (5)$$

TABLE 1. Parameters of feature space.

Parameters	Description	Datatype	Measure
distance between ourselves and the enemy	Distance between our battle units and enemy target.	numerical	km
speed of our battle units	Absolute velocity of our battle units.	numerical	km/h
altitude of our battle units	Height of our battle units.	numerical	km
type of enemy target's radiation source	Includes 3 types: radar, communication, jamming.	non-numerical	3-bit multi-hot encoding
working mode of the radiation source	Includes 9 working modes:Radar: search, track, fire control; Jamming: barrage noise, dense false target, smart noise, RVGPO; Communications: transmit, search.	non-numerical	9-bit multi-hot encoding

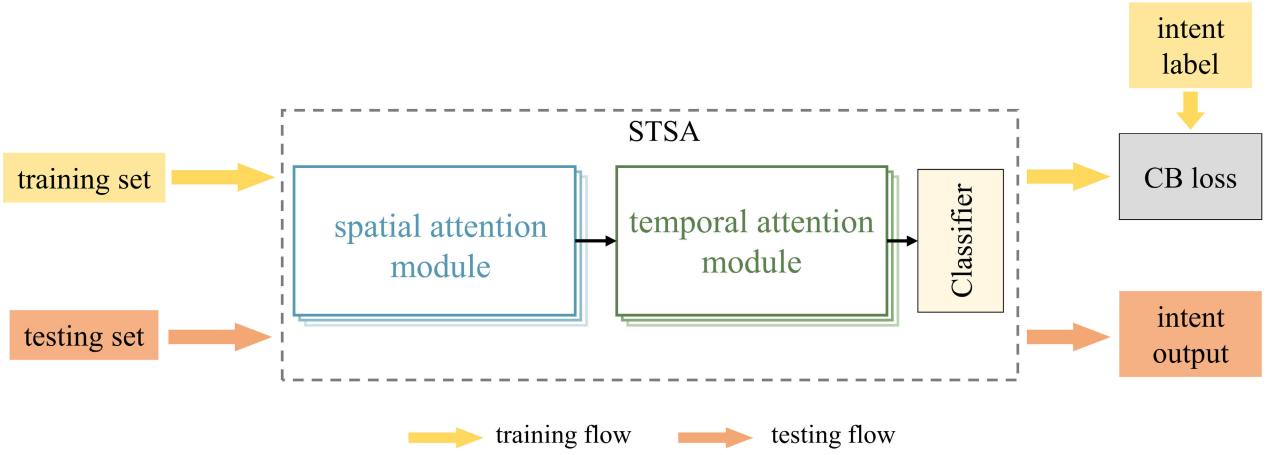

FIGURE 2. The overview of proposed CBSTSA model for combat intention recognition.

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ denote the query vector, key vector, and value vector, respectively. These vectors are obtained from the input $\boldsymbol{X}$ either through three linear layers, alternatively, by directly utilizing the input $\boldsymbol{X}$. d_k denotes the dimension of the query and key. The purpose of the self-attention mechanism is to recalibrate the input data by evaluating the correlation between each component of the data. Specifically, the attention term α of a particular component in relation to others is calculated by the scaled multiplicative attention, and then the value vectors are weighted by $\alpha \boldsymbol{V}$ to obtain the final self-attention output.

The multi-head self-attention mechanism, as the name suggests, uses multiple self-attention modules and combines the results of each attention module together as outputs by means of concatenation. The inputs to each attention head are obtained by linearly mapping the original $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ n times, and each attention head computes the outputs in parallel. All the results are concatenated together and passed through a linear layer to obtain the intention output. Assuming that $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{m \times d_{input}}$ are the inputs, the multi-head self-attention mechanism can be formally represented as:

$$\begin{aligned} &MultiHead\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) \\ &= Concat\left(head_1, head_2, \ldots, head_n\right) \boldsymbol{W}^O \end{aligned}$$

$$head_i = Attention\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \tag{6}$$

where $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K \in \mathbb{R}^{d_{input} \times d_k}$ and $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{input} \times d_v}$ denote the input mapping matrices. $\boldsymbol{W}^O \in \mathbb{R}^{nd_v \times d_{output}}$ denotes the output linear layer weight matrix of the multi-head attention. *Concat* denotes the concatenate operation.

C. CLASS BALANCED LOSS

As mentioned in Section I, the combat intention datasets exhibit a long-tailed distribution [12], with different intentions having varying importance. The presence of a long-tailed distribution in the dataset reflects the inherent asymmetry of real-world combat scenarios, which poses challenges for data-driven models, particularly for the recognition of the tail classes. These tail classes typically represent intentions with higher threat levels.

In this paper, we propose a novel CB loss function to address the issue of long-tailed distributions. The CB loss function operates by adaptively reweighting the traditional cross entropy (CE) loss function. Specifically, the reweighting term applied to the head classes is smaller compared to that applied to the tail classes, so that the model can allocate more attention to the tail classes. This adjustment ensures

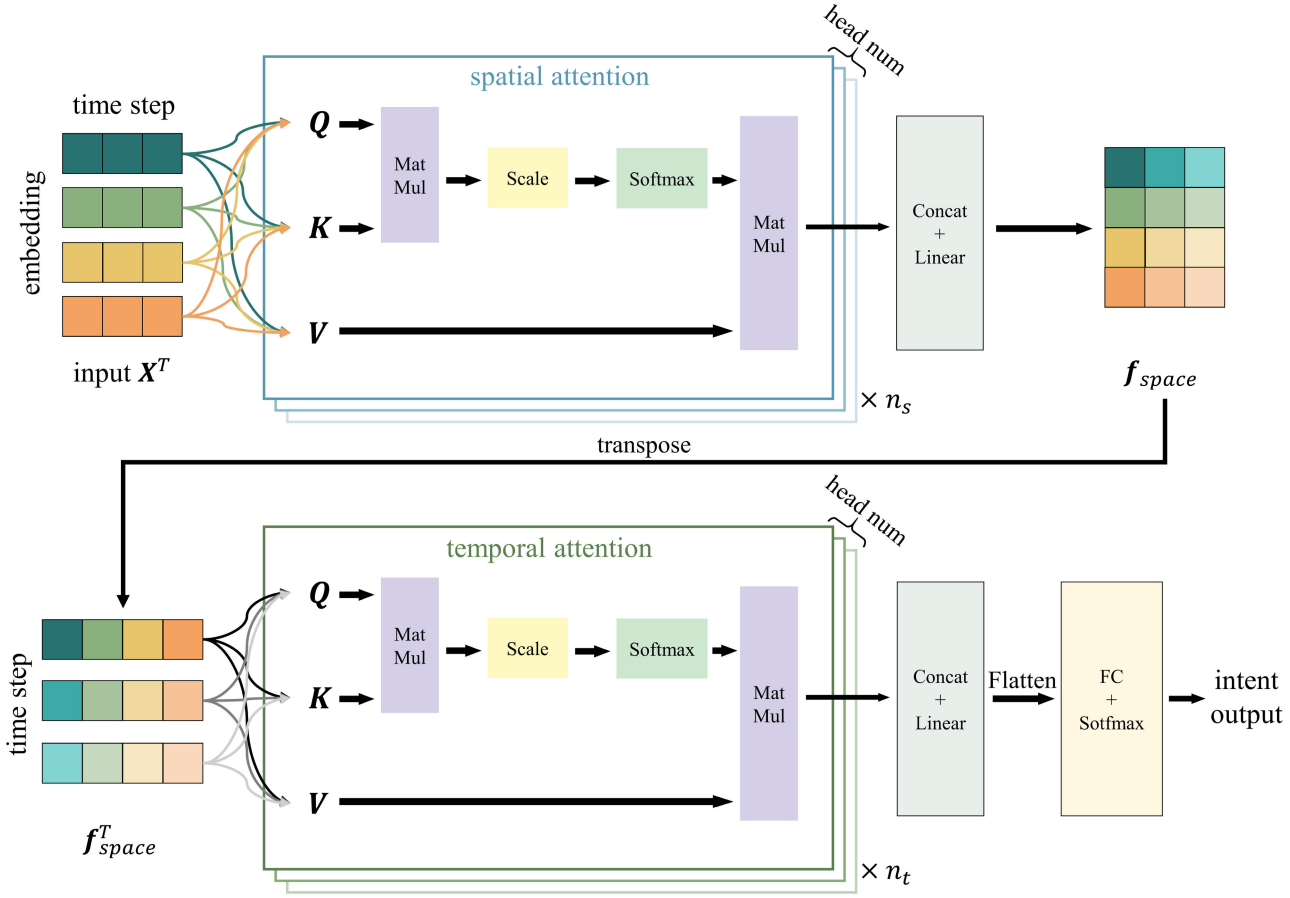

FIGURE 3. Structure of the spatio-temporal self-attention mechanism.

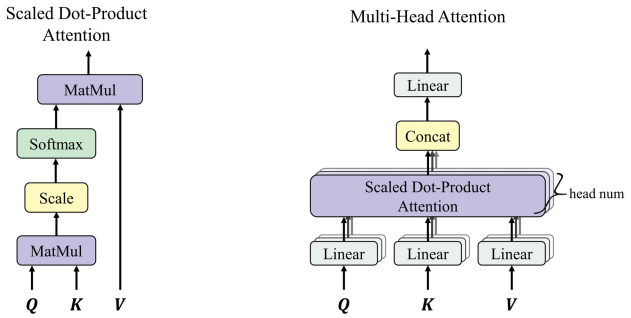

FIGURE 4. Self-attention mechanism unit.

that the abundance of samples in the head classes does not overshadow the importance of the tail classes.

The reweighting term in the CB loss function is designed by incorporating both the threat level of each intention and the effective number of samples. The threat level of each intention L_y is assumed as shown in Table 2.

The effective number of samples can be expressed as [24]:

$$\begin{cases} E_y = \frac{1-\beta^{n_y}}{1-\beta} \\ \beta = \frac{N-1}{N} \end{cases} \tag{7}$$

TABLE 2. Threat level of intention.

Intention	Threat Level L_y
Scout	1
pre-Intercept	2
Intercept	3
Obstruct	4
Lure	5

where E_y represents the effective number of samples for class y. n_y denotes the number of samples for class y. N denotes the volume of the sample space of the overall dataset, and usually $N \geq 1$. It is evident that the β is less than 1. When the number of samples for a particular class n_y is larger, the effective number of samples E_y will also be greater. This relationship is logical and aligns with our expectations.

Considering that the bigger E_y is, the greater penalty should be, we use $\alpha_i \propto L_i/E_i$ as the reweighting term of the class i. Formally, the raw cross entropy loss function for one sample is:

$$CE(\boldsymbol{p}, \boldsymbol{y}) = -\sum_{i=0}^{C-1} y_i \log(p_i) \tag{8}$$

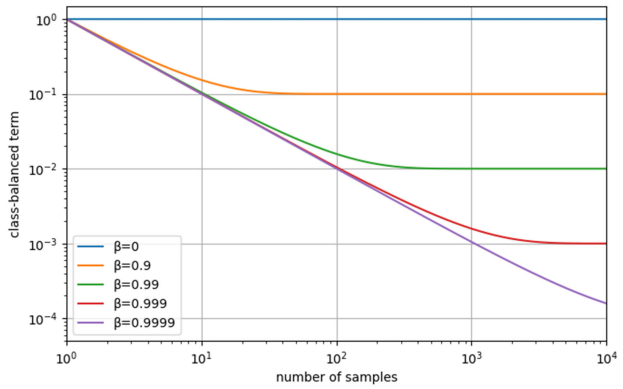

FIGURE 5. Relationship between class balanced term and sample size ($L_i = 1$).

where $\boldsymbol{p} = [p_1, p_2, \ldots, p_{C-1}]$ denotes the model prediction result, $p_i \in (0, 1)$ and $\sum_{i=1}^{C-1} p_i = 1$. $\boldsymbol{y} = [y_1, y_2, \ldots, y_{C-1}]$ denotes the label of the sample. $y_i = 1$, if the sample belongs to the class i, otherwise $y_i = 0$. C denotes the number of classes.

Further, reweighting CE loss using α_i, the CB loss function is given by:

$$CB(\boldsymbol{p}, \boldsymbol{y}) = -\sum_{i=0}^{C-1} \frac{(1-\beta)L_i}{1-\beta^{n_{y_i}}} y_i \log(p_i) \tag{9}$$

The CB loss can ensure the model focuses on recognizing and addressing the most critical scenarios. By assigning appropriate weights to different classes based on their threat levels and effective number of samples, the head classes will not dominate the training process and overshadow the importance of classes with fewer samples but higher threat levels.

In practice, since N is difficult to determine, this paper directly chooses β as the hyper-parameter, and adjusts the reweighting term by adjusting β. Different β will make the loss function have different sensitivity to the number of samples, and the closer β is to 1, the greater the penalty of the reweighting term for the head category samples will be, as shown in Fig. 5. In addition, we apply normalization to the reweighting term α_i during the experiments, so that the sum of all category reweighting terms equals the number of categories C.

$$\sum_{i=0}^{C-1} \alpha_i = C \tag{10}$$

The normalization step helps to ensure the scale of the loss function, thereby minimizing the impact of scaling variations in the loss values on the effectiveness of training process.

V. EXPERIMENTS

A. DATASET AND EXPERIMENTAL ENVIRONMENT

1) DATASET AND EVALUATION

The raw data comes from a combat simulation system, and we first use the sliding window to segment the raw time-series data. Subsequently, we submit this segment data to warfare experts for accurate labeling. The dataset comprises a total of 12,000 samples, with each sample being a two-dimensional time-series containing 10 simulation moments. Each moment has 15-dimensional feature parameters. Hence, one sample can be denoted as $\boldsymbol{X} \in \mathbb{R}^{10\times15}$. The dataset is divided into a training set and a test set according to the ratio of 8:2. The training set has a total of 9,600 samples, of which 47.083% are scout intention, 41.875% are pre-interceptor intention, 5.417% are interceptor intention, 5.417% are obstruct intention, and 0.208% are lure intention. The test set has a total of 2,400 samples, and each type of intentions has 480 samples.

Furthermore, this work obtains multiple test sets with different difficulty levels by adding different degrees of noise, as a way to test the robustness of the model in a complex battlefield environment. Specifically, signal-to-noise ratios (SNR) of 10dB, 5dB, 0dB, -5dB, -10dB is added to the numerical parameters (following the method in [3]), and the noise with error rate of 15%, 20%, 25%, 30%, 35% is added to the non-numerical parameters.

Following the previous approach [3], [4], [5], In this paper, accuracy, recall, precision, and F1 score are used as evaluation metrics. Moreover, the average-accuracy of the model on test sets with different noise levels is used to evaluate the robustness.

2) EXPERIMENTAL ENVIRONMENT

We programed in python 3.9 on 64 bit Windows 10 computer with the Tensorflow deep learning framework, a 12th Gen Intel(R) Core(TM) i7-12700 @ 2.10 GHz processor, and 32GB of RAM.

B. IMPLEMENTATION DETAILS

To determine the structure of the model, we need to set following parameters: the number of layers n_s for the multi-head attention in the spatial attention module, the number of layers n_t for the multi-head attention in the temporal attention module, and the β of the class balanced loss function. In order to make the model more attentive to the long-term temporal dependencies, we impose a constraint that the number of layers in spatial attention module not exceed that in the temporal attention modules, i.e., $n_s \leq n_t$. We further conduct a parameter search within $(n_s, n_t) \in (1, 1), (1, 2), (1, 4), (2, 4)$. For β, we perform parameter search among $\beta \in \{0, 0.9, 0.99, 0.999, 0.9999\}$, where $\beta = 0$ indicates that the standard cross entropy loss function is used for training. Additionally, the multi-head self-attention mechanism in the temporal and spatial attention module are configured with the following parameters: the number of attention heads h is set to 8, the dropout rate is set to 0.1, and $d_k = d_v = 64$.

The model is trained with each of the above configurations using the Adam optimizer, the learning rate is set to 0.001, over 50 epochs with a batch size of 200, and then evaluated on the original test set. As depicted in Fig. 6, the performance

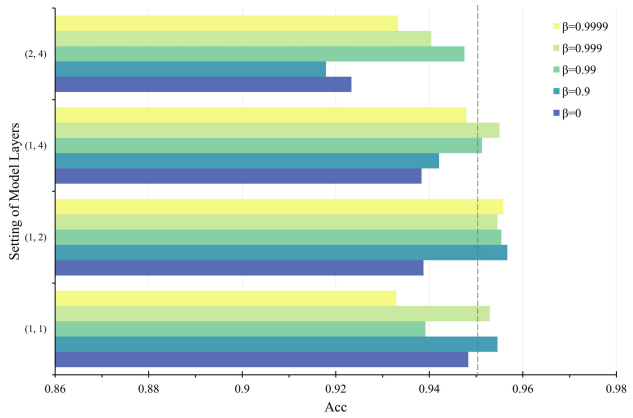

FIGURE 6. Performance of different model structures.

TABLE 3. Implementation details.

Parameters	Value
n_s	1
n_t	2
h	8
d_k	64
d_v	64
β	0.9
Dropout rate	0.1
Batch Size	200
Epoch	50
Optimizer	Adam
Learning rate	0.001

of the model significantly deteriorates compared to the other three configurations when the number of layers is set to $(2, 4)$. An excessive number of layers enables the model to extract more intricate features but also leads to overfitting. The complex patterns and features in the input data are not adequately extracted when the number of layers is set to $(1, 1)$, which leads to a weak effect of the class balanced loss function. The insufficient depth of feature extraction hampers the model's ability to capture and understand the intricate characteristics embedded within the data. Both the $(1, 2)$ and $(1, 4)$ configurations demonstrate optimal results exceeding 95% accuracy. However, the $(1, 2)$ configuration shows a more consistent and stable performance improvement. Considering training efficiency and stability, we determine the model structure as follows: $n_s = 1$, $n_t = 2$ and $\beta = 0.9$. The specific implementation details are shown in Table 3.

C. COMPARATIVE EXPERIMENT

To evaluate the performance of the proposed CBSTSA model in the combat intention recognition task, comparative experiments are designed. The CBSTSA model is evaluated by compared with several baseline models, including the MLP model, SAE model [6] GRU-FCN model [3], the 1DCNN-BiLSTM model [4], the TCN-attention model [5], and the Transformer encoder model. The structural parameters of all the aforementioned models are set according to the original papers. Note that the transformer encoder model uses the complete encoder module from [25]and stacks 3 layers, which aligns with the number of layers in CBSTSA model. And MLP is the basic neural network, which is also set to three layers, with 512, 256, and 128 neurons, respectively. The training regime is consistent across all models and is detailed in Section V-B. To ensure the fair and avoid contingency, each model undergoes training and testing processes five times, and the optimal results are selected for comparison. As shown in Table 4, the CBSTSA model achieves an accuracy of 95.67%, showcasing a significant performance improvement over the baseline models. Specifically, it outperforms the GRU-FCN, 1DCNN-BiLSTM, TCN-attention, and Transformer encoder by margins of 5.04%, 5.29%, 4.25%, and 1.54%, respectively.

TABLE 4. Results of original test dataset. Acc is an abbreviation for accuracy. Precision, Recall, and F1 scores are all macro-averaged.

Model	Acc (%)	Precision	Recall	F1
MLP	89.88	0.9080	0.8988	0.8981
SAE	89.75	0.9149	0.8975	0.8964
GRU-FCN	90.63	0.9241	0.9063	0.9048
1DCNN-BiLSTM	90.38	0.9220	0.9037	0.9021
TCN-attention	91.42	0.9265	0.9142	0.9135
Transformer encoder	94.13	0.9460	0.9412	0.9414
CBSTSA	**95.67**	**0.9587**	**0.9567**	**0.9568**

Further, to validate the robustness of the model, an additional comparative experiment is conducted on test sets with different levels of noise, as indicated by the signal-to-noise ratio (SNR). It is crucial to noted that, this work refrains from retraining the model under the different SNR. Instead, the model trained on the original dataset is directly used for testing with different noise, which is different from the method in [3]. By adopting this methodology, we aim to evaluate the model's generalization and adaptability to noisy environments, which aligns with real-world conditions. Results in Table 5 shows that our model consistently maintains optimal performance across all SNR, exhibiting robustness to varying levels of noise. On average, it achieves an accuracy improvement of 6.08%, 6.57%, 4.35%, 6.88%, 3.72%, and 2.22% compared to the other models, respectively.

In addition, we calculate the gradient of the linear trend line to quantify the model's susceptibility to noise, offering a quantitative measure of its sensitivity and further elucidating its performance profile. As depicted in Fig. 7, the calculated gradients are -3.28 for TCN-attention, -4.04 for the Transformer encoder, and -3.31 for our proposed model. It means that as the SNR decreases, our model exhibits a slower or similar performance decline compared to the other models. This suggests that our model displays greater resilience to noise and enhanced adaptability to the complexities of the battlefield environment.

TABLE 5. Results of robustness test. We compare the accuracy of the models at different SNRs. The SNR is set to /, 10, 5, 0, -5, -10 in dB, where / stands for the original test dataset.

SNR (dB)	Model	/	10	5	0	-5	-10	Avg.
Acc (%)	MLP	89.88	88.08	86.21	82.92	79.33	69.79	82.70
	SAE	89.75	88.13	85.38	82.42	79.38	68.21	82.21
	GRU-FCN	90.63	89.46	87.75	84.54	82.08	72.13	84.43
	1DCNN-BiLSTM	90.38	89.75	87.21	85.00	80.63	58.46	81.90
	TCN-attention	91.42	89.96	87.71	85.17	82.92	73.21	85.06
	Transformer encoder	94.13	92.21	90.38	87.25	84.00	71.38	86.56
	CBSTSA	**95.67**	**92.96**	**91.46**	**89.54**	**86.00**	**77.08**	**88.78**

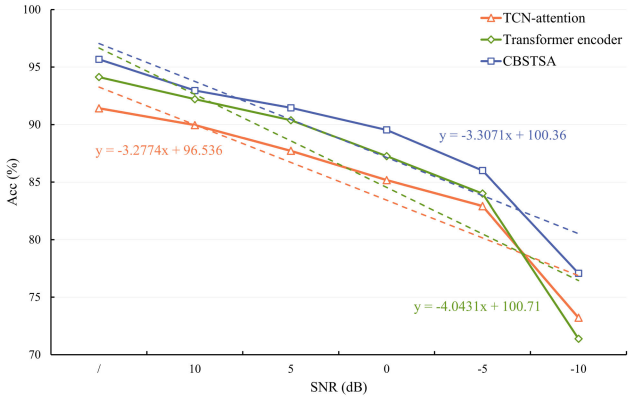

FIGURE 7. Linear trendlines for baseline models and CBSTSA. The dashed line represents downward trend in model performance.

Considering the requirements of lightweight and timely for intention recognition models in real combat scenarios, we count the parameter quantities and FLOPs for above models to quantify the model's size and computational complexity. As shown in Table 6, although MLP and SAE have lower computational demands, their performance is suboptimal. The FLOPs of other models are in the million range, but our CBSTSA model is more lightweight, with only 87k parameters. Therefore, considering the model size, computational complexity, and performance, CBSTSA is the optimal choice.

D. ABLATION EXPERIMENT

To evaluate the influences of each component, we design ablation experiments for the class balanced loss function, the spatial attention module, and the temporal attention module. Note that the model is trained based on the original cross entropy loss function when the class balanced loss function is removed. Table 7 shows the effectiveness of spatial-temporal structure. The spatio-temporal self-attention model can achieve the best performance when same loss function is used for training. Especially, when the class

TABLE 6. Statistics on model size and computational complexity. Params represents the number of model parameters, and FLOPs represents the number of floating-point operations a model has for one sample.

Model	Params	FLOPs
MLP	242k	0.48M
SAE	84k	0.18M
GRU-FCN	281k	5.55M
1DCNN-BiLSTM	110k	13.8M
TCN-attention	203k	1.68M
Transformer encoder	103k	4.08M
CBSTSA	87k	4.64M

TABLE 7. Results of ablation experiments. ✓ indicates that the component is activated. When the Class Balanced loss is activated, we annotate the values of the hyperparameter β.

Component			Acc (%)	Avg. (%)[a]
Spatial Attention	Temporal Attention	Class Balanced loss		
✓			92.96	85.92
✓		✓($\beta = 0.99$)	94.25	86.97
	✓		93.96	86.15
	✓	✓($\beta = 0.9999$)	94.50	85.13
✓	✓		94.29	86.37
✓	✓	✓	**95.67**	**88.78**

[a] Avg. denotes the average accuracy of models under different SNRs.

balanced loss is employed, the Acc (Avg.) of STSA can be improved by 1.42% (1.81%) and 1.17% (3.65%) compared to the spatial-attention based model and the temporal-attention based model, respectively.

Furthermore, the class balanced loss function can improve the model performance to a certain degree. Specifically, 1.29% for the spatial-attention based model, 0.54% for the temporal-attention based model, and 1.38% for the STSA model. As shown in Table 8, we also compare the F-1 scores (based on STSA model) for both head and tail classes with and without the CB loss. The results suggest that the CB loss function can improve the performance of tail intentions without compromising the performance of head one, thus achieving an improvement in overall performance. More intuitive display of tail performance improvement is shown in Fig. 8.

E. RESULT DISCUSSION

The extensive experiments demonstrate the solid improvement of our model in the combat intent recognition task. This improvement is largely attributed to the STSA mechanism and CB loss function. More specifically, in the comparative experiments with other baseline models (as shown in Table 4), it can be found that Transformer encoder model [25] and our CBSTSA model achieve the top 2 accuracy. We consider the capacity of the self-attention mechanism to capture contextual features in sequential data as the crucial factor

TABLE 8. F-1 scores for head and tail classes with and without CB Loss. CE stands for the cross-entropy loss. Scout is the head class, and lure is the tail class.

Molde		STSA			
Loss		CE		CB	
Class		scout	lure	scout	lure
SNR	/	**0.9885**	0.9154	0.9874	**0.9466**
	10dB	**0.9875**	0.8832	0.9832	**0.9186**
	5dB	0.9769	0.8493	**0.9810**	**0.8902**
	0dB	**0.9708**	0.8287	0.9703	**0.8632**
	-5dB	0.9499	0.8152	**0.9522**	**0.8427**
	-10dB	0.8441	0.6900	**0.9028**	**0.8045**

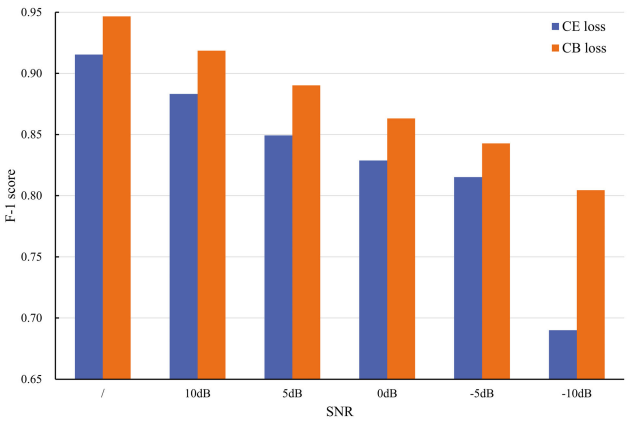

FIGURE 8. Illustration of the F-1 score for tail class (lure intention).

underlying this phenomenon. However, as the level of noise increases, the advantages of the CBSTSA model become more pronounced (as shown in Table 5), and we attribute this to the effect of the spatial attention module. To further validate this opinion, we conduct ablation experiment, and Table 7 illustrates that when spatial and temporal attention modules are simultaneously activated, the performance is superior (no matter with or without CB loss). Furthermore, we conduct validation on the effect of the CB loss, as presented in Table 8 and Fig. 8. The incorporation of the CB loss enables the model to achieve a significant improvement of approximately 4% on tail classes, while maintaining the performance on head classes.

VI. CONCLUSION

A novel CBSTSA network for combat intention recognition is proposed. By implementing the self-attention mechanism in the feature space dimension, the STSA enhances the feature of the original time series data and effectively improves the feature expressions for classification. Additionally, by incorporating the sample size and threat level into the reweighting term, the CB loss function can effectively deal with the inherent asymmetry in battlefield intention data. Extensive experiments show that both the spatial attention module and CB loss are pivotal in advancing combat intention recognition. In future work, we try to extend the CBSTSA framework to more combat scenarios.

REFERENCES

[1] D. L. Wang, X. F. Wu, and H. P. Leng, "Some problems for intention assessment to foe in battle-field," *Ship Electron. Eng.*, vol. 24, no. 6, pp. 4–9, Jun. 2004.
[2] H. W. Wang, H. Q. Shi, and X. Z. Zhao, "A summary of target intention identification methods," in *Proc. China High-Level Forum Syst. Simulation VR Technol.*, Beijing, China, 2020, pp. 186–188.
[3] P. Ding and Y. F. Song, "A cost-sensitive method for aerial target intention recognition," *Acta Aeronaut. et Astronaut. Sin.*, vol. 45, no. 2, pp. 176–191, 2024.
[4] C. Zhang, Y. Zhou, H. Li, F. Liang, Z. Song, and K. Yuan, "Combat intention recognition of air targets based on 1DCNN-BiLSTM," *IEEE Access*, vol. 11, pp. 134504–134516, 2023.
[5] L. Zhao, P. Sun, and Y. J. Zhang, "A fast aerial targets intention recognition method under imbalanced hard-sample," *J. AF Eng. Univ.*, vol. 25, no. 1, pp. 76–82, Jan. 2024.
[6] W. Ou, S. J. Liu, X. Y. He, and S. M. Guo, "Tactical intention recognition algorithm based on encoded temporal features," *Command Control Simul.*, vol. 38, no. 6, pp. 36–41, Dec. 2016.
[7] X. Liu, M. Zhao, S. Dai, Q. Yin, and W. Ni, "Tactical intention recognition in wargame," in *Proc. IEEE 6th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Chengdu, China, Apr. 2021, pp. 429–434.
[8] Y. Li, J. Wu, W. Li, W. Dong, and A. Fang, "A hierarchical aggregation model for combat intention recognition," *J. Northwestern Polytech. Univ.*, vol. 41, no. 2, pp. 400–408, Apr. 2023.
[9] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554–2558, Apr. 1982.
[10] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.
[11] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.
[12] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Shinjuku, Japan, 2017, pp. 115–124.
[13] J. Azarewicz, G. Fala, and C. Heithecker, "Template-based multi-agent plan recognition for tactical situation assessment," in *Proc. Conf. Artif. Intell. Appl.*, Miami, FL, USA, 1989, pp. 247–254.
[14] X. Xia, "The study of target intent assessment method based on the template-matching," M.S. thesis, Dept. Autom, Graduate School NUTD, Changsha, China, 2007.
[15] Y. Wang, "Research on battlefield target identification and situation intention forecasting," M.S. thesis, Dept. Comput. Sci., Jiangnan Univ., Wuxi, China, 2015.
[16] Y. Xiao-hong and C. Jia-yuan, "An improved template-matching based communication object recognition," *Mod. Def. Technol.*, vol. 46, no. 5, p. 69, Oct. 2018.
[17] V. P. Kirillov, "Constructive stochastic temporal reasoning in situation assessment," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 8, pp. 1099–1113, Aug. 1994.
[18] F. J. Zhao, Z. J. Zhou, C. H. Hu, L. Wang, and T. Liu, "Aerial target intention recognition approach based on belief-rule-base and evidential reasoning," *Electron. Opt. Control*, vol. 24, no. 8, pp. 15–19, Aug. 2017.
[19] D. Kong, T. Chang, Q. Wang, H. Sun, and W. Dai, "A threat assessment method of group targets based on interval-valued intuitionistic fuzzy multi-attribute group decision-making," *Appl. Soft Comput.*, vol. 67, pp. 350–369, Jun. 2018.
[20] H. Deng, Q. J. Yin, J. Hu, and Y. B. Zha, "Tactical intention recognition based on multi-entity Bayesian network," *Syst. Eng. Electron.*, vol. 32, no. 11, pp. 2374–2379, Nov. 2010.
[21] Z. X. Yu, X. X. Hu, and W. Xia, "Foe intention inference in air combat based on fuzzy dynamic Bayesian network," *J. HeFei Univ. Technol.*, vol. 36, no. 10, pp. 1210–1216, Oct. 2013.
[22] Y. L. Sun and L. Bao, "Study on recognition technique of targets' tactical intentions in sea battlefield based on D-S evidence theory," *Ship Electron. Eng.*, vol. 32, no. 5, pp. 48–51, May 2012.

[23] F. J. Baldán and J. M. Benítez, "Multivariate times series classification through an interpretable representation," *Inf. Sci.*, vol. 569, pp. 596–614, Aug. 2021.
[24] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9260–9269.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkorei, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 1–11.

XUAN WANG received the B.E. degree from Tianjin University (TJU), Tianjin, China, in 2022. He is currently pursuing the M.E. degree in electronic information with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests mainly include combat intention recognition and knowledge graph.

BENZHOU JIN (Member, IEEE) received the B.E. degree in instrument science and technology from Xidian University, Xi'an, in 2008, and the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, in 2013. He is currently a Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing. From 2013 to 2018, he was a Senior Engineer with Nanjing Research Institute of Electronics Technology, Nanjing. From 2018 to 2022, he was an Associate Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics. His current research interests include radar system engineering, nonlinear signal processing, and array signal processing.

MINGYANG JIA received the B.E. degree from Henan University of Science and Technology (HAUST), Henan, China, in 2021. He is currently pursuing the M.E. degree in electronic information with the College of Electronic and Information Engineering, Electronic Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests mainly include aliased signal separation and main lobe interference suppression of radar systems.

GANG WU received the B.E. degree in electrical engineering from Xinjiang University, Xinjiang, in 1995, and the M.S. degree in mechanical engineering from Huazhong University of Science and Technology, Wuhan, in 2001. He is currently a Professor with 14th Research Institute, China Electronics Technology Group Corporation, Nanjing, China. His current research interests include space-time adaptive processing and array signal processing.

XIAOFEI ZHANG received the M.S. degree from Wuhan University, Wuhan, China, in 2001, and the Ph.D. degree in communication and information systems from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2005. He is currently a Full Professor with the Electronic Engineering Department, Nanjing University of Aeronautics and Astronautics. His research interests include array signal processing and communication signal processing. He is on the Technical Program Committees of the IEEE 2010 International Conference on Wireless Communications and Signal Processing, the IEEE 2011 International Conference on Wireless Communications and Signal Processing, SSME2010, and 2011 International Workshop on Computation Theory and Information Technology. He is an Editor of *International Journal of Digital Content Technology and Its Applications*, *International Journal of Technology and Applied Science*, *Journal of Communications and Information Sciences*, *Scientific Journal of Microelectronics*, and *International Journal of Information Engineering*.