



OPEN Interpretable capsule networks via self attention routing on spatially invariant feature surfaces

Peizhang Li^{1,3}, Jiyuan Ru^{1,3}, Qing Fei¹✉, Zhen Chen¹ & Bo Wang²

The accurate and efficient evaluation and classification of situational images is fundamental to making informed and effective decisions. However, current classification approaches based on convolutional neural networks often suffer from limited generalization and robustness, particularly when processing data characterized by abstract class features and pronounced spatial attributes. Additionally, the “black-box” nature of deep neural network architectures poses significant challenges to their application in fields with stringent security requirements. To address these limitations, this paper introduces a novel Spatially Invariant Self-Attention Capsule Network (SISA-CapsNet), designed to encode interpretable spatial features for classification tasks. SISA-CapsNet employs capsules to encode spatial features from specific image regions and classifies these features through a self-attention routing mechanism. Specifically, spatially invariant feature surfaces with dimensions identical to the input image are generated and stacked to form feature capsules, each encoding spatial features from distinct regions. The self-attention mechanism calculates coupling coefficients, clustering feature capsules into class capsules. This architecture integrates a spatially invariant feature extraction structure, facilitating pixel-level encoding of regional spatial features, and leverages self-attention to effectively capture the relative importance of different spatial regions for classification. Together, these two mechanisms constitute an interpretable classification framework. Experimental validation on benchmark datasets and battlefield situational image datasets with pronounced spatial characteristics demonstrates that the proposed method not only achieves superior classification performance but also offers interpretability closely aligned with human cognitive processes. Furthermore, comparative analyses with existing visual interpretability methods underscore the enhanced interpretability of SISA-CapsNet.

Keywords Situational image classification, Interpretability, Capsule network, Self-attention

Situation assessment involves the extraction of features, recognition of patterns, and estimation of intent from observational data to construct a comprehensive situational image. This process is widely utilized in domains such as command and control decision-making, traffic management, and cybersecurity¹. The accurate differentiation of situational images across diverse categories is essential for enabling rapid, stable, and effective decision-making². Computer vision has emerged as a key technology for addressing this challenge. Traditional computer vision methods rely on feature extraction operators within classifiers; however, their performance is often limited by environmental factors such as viewpoint variation, object posture, and occlusion³, leading to reduced robustness and generalization. Moreover, semantic features, which include abstract concepts and spatial relationships, are challenging to capture using feature operators⁴, thereby restricting the applicability of traditional methods to tasks like situational image classification. Recently, artificial intelligence, particularly deep learning-based approaches, has demonstrated superior feature extraction and nonlinear fitting capabilities^{5–7}, offering effective solutions to these limitations^{8–12}.

The remarkable capabilities of deep learning arise from its complex nonlinear structures and numerous parameters, resulting in an opaque “black-box” architecture^{13,14}. This lack of interpretability hinders understanding of its operational principles and reasoning mechanisms, significantly limiting its application in situational image classification tasks within domains requiring high security and sensitivity. In response to this challenge, explainable artificial intelligence (XAI) techniques have gained increasing attention and development. Existing XAI methods can be broadly categorized into three types, model self-explanation, specific model

¹School of Automation, Beijing Institute of Technology, Beijing 100081, China. ²China Shipbuilding Zhihai Innovation Research Institute, Beijing, China. ³These Authors contributed equally: Peizhang Li and Jiyuan Ru ✉email: feiqing@bit.edu.cn

explanation, and model-agnostic explanation. Model self-explanation methods, such as linear regression¹⁵ and decision trees¹⁶, build transparent linear networks in advance to achieve interpretable classification. However, the linear nature of these models limits their ability to capture the complex nonlinear features inherent in situational image classification, resulting in suboptimal performance in both classification and generalization. Specific model explanation methods involve visualizing the internal structure and parameters of neural networks to interpret decision-making, as seen in techniques like Class Activation Mapping (CAM)¹⁷, Grad-CAM¹⁸, and Activation Sensitivity¹⁹. Although these methods allow for the direct visualization of gradient information, they do not capture the spatial semantic information critical to situational images. Model-agnostic explanation methods, such as LIME²⁰, Kernel SHAP²¹, and knowledge distillation²², provide indirect explanations based on input-output relations. These methods are limited in that they offer only local approximations and fail to provide global explanations of the model's decision-making process, making it difficult to generate the trustworthy neural networks desired for such tasks. In the context of situational image classification, interpretability is particularly critical for the following reasons, (a) Situational image classification requires not only accurate predictions but also transparency in the model's decision-making process and causal reasoning to prevent scenarios where correct results arise from flawed reasoning. (b) Effective situational image classification involves human-machine collaboration, where machines must incorporate expert knowledge and humans rely on machines for data processing. Limited interpretability hampers the synergy between human and machine intelligence. (c) Situational image classification tasks demand models that can extract interpretable, semantically meaningful spatial features from high-noise, small-sample datasets, ensuring robust and generalizable performance in complex, dynamic environments. In light of these limitations, our objective in the context of situational image classification is to develop a transparent classification model that encodes spatial features interpretably, thereby transforming the decision-making process into a linear combination of these features.

Capsules, a novel neural unit introduced by Hinton^{23,24}, present an alternative to traditional neural units and can be utilized to construct explainable capsule network models. Each capsule encodes the instantiation parameters of a specific entity type as an activity vector, where the vector's length reflects the probability of the entity's occurrence²⁵. This vectorized feature representation, along with the linear combination of capsules, fulfills the requirements for addressing the interpretability challenges discussed in this paper. However, existing research on capsule networks predominantly concentrates on optimizing the routing mechanism to enhance performance^{26–28}, with relatively little focus on their interpretability^{29–31}, especially regarding spatial feature interpretability. Current studies have primarily explored the use of capsule vectors to encode features, yet they have not addressed how these features correspond to interpretable spatial information. Achieving intuitive interpretability hinges on constraining the sources of features encoded by capsules at the spatial level, ensuring that these encodings possess clear spatial semantics that can be directly matched with the input image.

Based on this, this paper proposes the Spatial Invariant Self-Attention Capsule Network (SISA-CapsNet), which visually provides the interpretable classification features in situational image classification, as shown in Fig. 1. SISA-CapsNet constructs feature surfaces that match the dimensions of the original input image, stacking feature surfaces of varying dimensions to form feature capsules. A self-attention mechanism is applied between feature capsules to compute the routing coefficients relative to the class capsules, which are then linearly combined to generate the class capsules. Furthermore, the spatial feature activation weight mask, derived from the routing coefficients, enables intuitive visualization of the influence of different spatial feature capsules, thereby facilitating visual interpretability. To evaluate the accuracy and interpretability of SISA-CapsNet, a series of controlled experiments were conducted across multiple datasets. First, the MNIST dataset was used to assess the basic performance of SISA-CapsNet. Next, the MSTAR dataset was employed to examine the model's classification performance and interpretability in scenarios with moderate spatial feature dependence and noise. Finally, a situational image classification dataset was utilized to test SISA-CapsNet's performance in tasks involving semantic abstraction and strong spatial feature dependence. The results demonstrate that SISA-CapsNet not only excels in classification tasks but also provides a classification rationale that aligns with human understanding. The contributions of this paper can be summarized as follows:

- We propose a novel interpretable image classification network, SISA-CapsNet, which provides intuitive visualizations of the primary spatial features leveraged by the neural network during image classification.
- SISA-CapsNet introduces an innovative integration of spatially invariant convolution to constrain feature capsules to encode specific spatial regions and incorporates a capsule self-attention mechanism to facilitate the linear combination of feature capsules. This design establishes a linear, transparent, and interpretable classification pathway.
- Additionally, we develop a novel spatial feature activation weight mask to enhance the visualization of SISA-CapsNet's interpretability. Comparative evaluations against existing image classification algorithms and interpretable classification methods demonstrate the effectiveness of the proposed approach.

The remainder of this paper is organised as follows. Section [Related Work](#) reviews related work pertinent to this paper. Section [Methods](#) provides a detailed explanation of the proposed method, including the SISA-CapsNet network architecture and its interpretability framework. Section [Experiments and Results](#) presents the datasets and experimental setup used for validation, showcasing both the classification performance and visual interpretability of SISA-CapsNet. Finally, Section [Conclusion](#) concludes the paper and discusses potential avenues for future research.

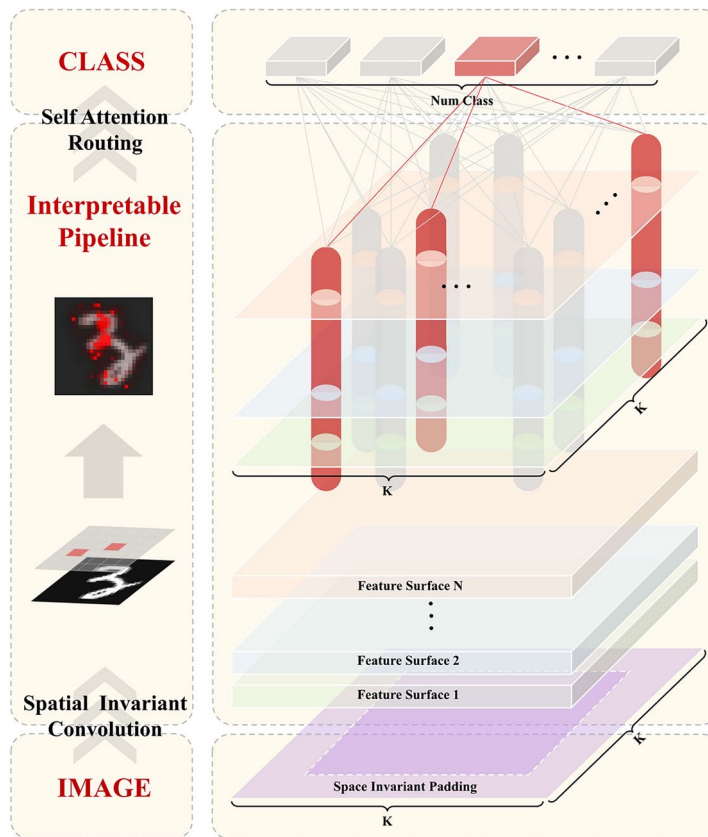


Fig. 1. The schematic framework of SISA-CapsNet.

Related work

Capsule network

Hinton et al.²³ introduced capsule networks, which replace traditional scalar neurons with vectorized neurons that encode specific entity types³². This approach effectively mitigates the loss of relative spatial information between features caused by pooling operations in convolutional neural networks (CNNs). However, the dynamic routing mechanism in capsule networks imposes constraints on both representational capacity and computational efficiency. To enhance the routing mechanism and improve training efficiency, Rajasegaran et al.²⁶ proposed a novel dynamic routing algorithm that integrates 3D convolutions to reduce the parameter count. Mazzia et al.²⁷ introduced a self-attention-based, non-iterative routing algorithm designed to improve both performance and computational efficiency. Gu et al.²⁸ developed Aff-CapsNets, which replace dynamic routing with a shared transformation matrix between all lower-level and higher-level capsules, thereby reducing computational resource consumption. Additionally, Liu et al.²⁹ incorporated attention mechanisms into pre-trained convolutional capsule networks, enhancing both recognition accuracy and efficiency in complex, variable tasks. The vectorized feature representation of capsule networks is also considered promising for improving interpretability within feature space, prompting further research in this domain. Huang et al.³⁰ introduced a dual-attention mechanism, integrating attention modules after convolutional and lower capsule layers to enhance image feature extraction, thereby improving the contribution of significant image features to capsule representations and increasing interpretability. Jung et al.³¹ proposed an interpretable capsule framework, iCaps, which addresses challenges such as classification-irrelevant information within capsules and entity overlap. Afshar et al.³³ demonstrated the interpretability of capsule networks in brain tumor image classification, while Sezavar et al.³⁴ developed DCapsNet, a deep neural network that employs five convolutional layers and a capsule network to capture local spatial features from sensor input data while preserving spatial relationships. This approach has yielded superior interpretability in human activity and gait recognition. Although these studies have contributed to advancing the interpretability of capsule networks, they primarily focus on generalized feature encoding. A key challenge remains in constraining the source of interpretable features directly from input data to enable end-to-end interpretability, an open problem that warrants further investigation.

Self-attention

The self-attention mechanism, introduced by Vaswani et al.³⁵, has demonstrated remarkable capabilities in modeling contextual dependencies^{36,37}, making it highly effective for capturing long-range dependencies in sequential data and explicitly representing intrinsic correlations among multidimensional features³⁸. Its global receptive field and dynamic feature selection mechanism have further facilitated its application in computer

vision^{39,40}. For example, Guo et al. proposed the Vision Attention Network (VAN)⁴¹, which is based on Large Kernel Attention (LKA) and effectively addresses the limitations of traditional self-attention mechanisms in image tasks. These limitations include the neglect of 2D structural information, high computational complexity, and a lack of channel adaptability. Cheng et al. introduced the Class Attention Network (CANet)⁴², which utilizes attention encoding learning to achieve class-aware attention encoding by disentangling feature entanglement. This approach overcomes the challenge of traditional attention mechanisms in explicitly extracting class-discriminative features. Wu et al. proposed the Dual-Branch Residual Attention Network (DRANet)⁴³, which employs parallel branches—namely, the Residual Attention Block (RAB) and the Hybrid Dilated Residual Attention Block (HDRAB)—to capture both global and local features while filtering out irrelevant information, thereby enhancing image denoising performance. Recently, the integration of capsule networks with self-attention mechanisms has emerged as a growing research trend, as self-attention enhances the feature encoding capability of capsules. For instance, Hoogi et al. incorporated a self-attention mechanism following the convolutional layers of a capsule network⁴⁴, improving global feature extraction and enhancing the integration of capsule feature encoding with salient image features, thereby boosting image classification performance. Similarly, Yu et al. introduced a self-attention mechanism into the feature extraction layers of capsule networks⁴⁵, improving feature representation and addressing the limitations of conventional convolutional networks. Ke et al. employed multi-head self-attention (MSA)⁴⁶ to derive richer feature representations from image data, while similar contributions have been made by Mazzia et al.²⁷ and Huang et al.³⁰. Despite these advancements, existing research on integrating capsule networks with self-attention has primarily focused on performance improvements and feature extraction optimization, with limited attention to interpretability. A key challenge that this study seeks to address is how to leverage self-attention to integrate pixel-level spatial features, thereby enabling end-to-end interpretability.

Methods

This section provides a detailed explanation of the SISA-CapsNet architecture proposed in this paper, along with a discussion of its interpretability. Furthermore, a visualization method is designed to intuitively demonstrate the model’s interpretability.

Spatially invariant self-attention capsule network

SISA-CapsNet comprises two primary components: the spatially invariant capsule generator (SI-Generator) module and the self-attention capsule routing (SA-Routing) module, as illustrated in Fig. 2. The network processes an input image through spatially invariant convolutional layers to generate feature maps, which are then stacked to form feature capsules. These feature capsules are weighted by the self-attention routing mechanism to generate class capsules. After training, the correct class capsule exhibits the maximum length, while the feature capsules are assigned normalized weights.

The SI-Generator module extracts features from the input image across multiple dimensions using convolution, compresses them, and stacks them to form feature vectors, referred to as capsules. In SISA-CapsNet, capsules are specifically constrained to encode spatial features of designated regions. They are generated by stacking multiple feature surfaces of the same size as the input image, as illustrated in Fig. 3.

To achieve this, the spatial invariant padding strategy is employed to pad an input image of size $K * K$. Under the condition that the stride is restricted to 1, the padding size SP is related to the size of the capsule encoding spatial region K_{Enc} ,

$$SP = \frac{K_{Enc} - 1}{2}. \tag{1}$$

The padded image is subjected to 2D convolution to generate a feature matrix with C channels. Feature capsules s_i are then formed by stacking the results of convolution and $1 * 1$ compressed convolutions,

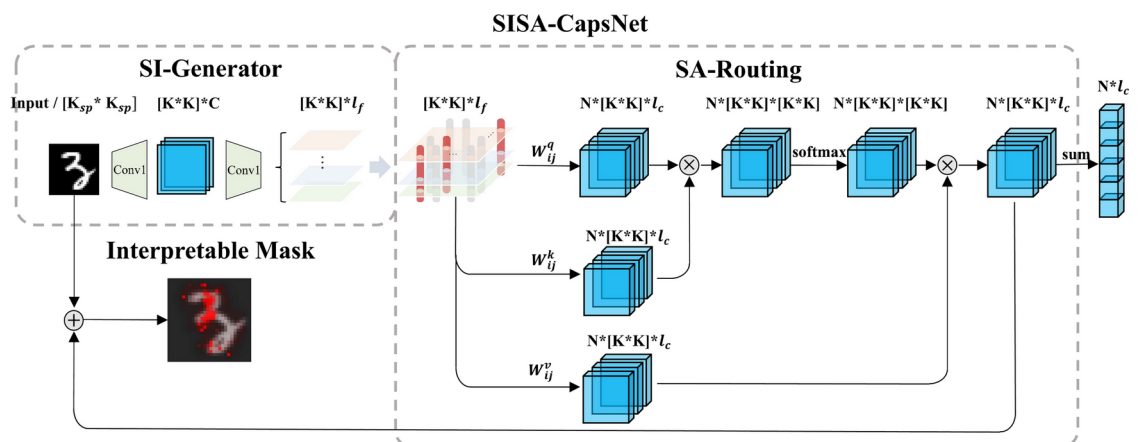


Fig. 2. The architecture of SISA-CapsNet algorithm.

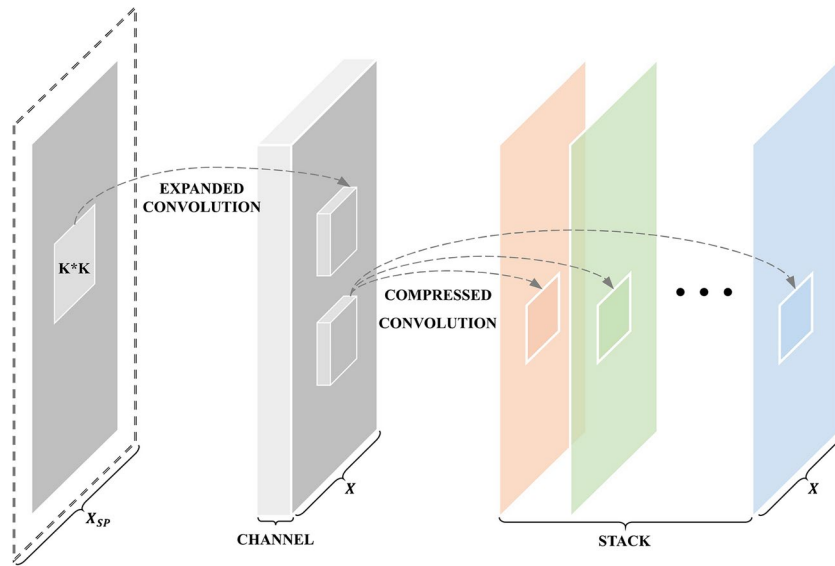


Fig. 3. The space invariant capsule extract module SI-Generator.

$$v_i = [conv^*(conv(X_{SP}))]^{l_f}. \tag{2}$$

Here, X_{SP} represents the input image after padding, $conv(\cdot)$ denotes feature extraction convolution, and $conv^*(\cdot)$ indicates compressed convolution. The parameter l_f refers to the number of stacking operations, which corresponds to both the number of feature surfaces and the length of the feature capsule s_i . Thus, in SI-Generator, spatial feature encoding is achieved through the first layer of spatial convolution, while the second layer of compressed convolution enables the stacking of features across different dimensions. This process ultimately produces multiple feature surfaces of the same sizes as the input image, which are then stacked to form spatially invariant feature capsules, with each capsule encoding features from a fixed spatial region.

In capsule networks, the linear relationship between lower-level feature capsules and higher-level class capsules is established through routing^{24,47}. Traditional iterative dynamic routing²⁴, which utilizes similarity-based inner product metrics, not only constrains the clustering capacity of capsules but also faces challenges in parallelization, limiting computational efficiency. Additionally, as the number of capsules increases, the receptive field for clustering in traditional routing diminishes significantly, preventing effective clustering on spatially invariant feature surfaces where the number of capsules corresponds to the number of pixels in the input image.

To address this issue, we propose the SA-Routing module that replaces traditional routing for the linear clustering of feature capsules into class capsules. The self-attention mechanism captures the correlations among all feature capsules on a feature surface, providing a robust global receptive field. For a feature capsule s_i (of length l_f), the capsule correlation prediction matrix q_{ij} and key prediction matrix k_{ij} for N class capsules s_j (of length l_c) are computed using transformation matrices W_{ij}^q and W_{ij}^k , respectively,

$$q_{ij} = s_i * W_{ij}^q [l_f * l_c], \tag{3}$$

$$k_{ij} = s_i * W_{ij}^k [l_f * l_c]. \tag{4}$$

The capsule correlation prediction matrix q_{ij} and the key prediction matrix k_{ij} are transposed, multiplied, and scaled. A softmax operation is then applied to generate the coupling coefficient matrix c_{ij} between the class capsules and feature capsules,

$$c_{ij} = softmax\left(\frac{q_{ij} * (k_{ij})^T}{\sqrt{d_k}}\right), \tag{5}$$

where d_k represents the scaling factor. The shape prediction matrix v_{ij} for feature capsule s_i with respect to the different class capsules s_j is computed,

$$v_{ij} = s_i * W_{ij}^v [l_f * l_c]. \tag{6}$$

The class capsules are derived by applying the coupling coefficient matrix and shape predictions matrix, which are then weighted and subsequently computed using the squashing function,

$$s_j = squash(c_{ij} * v_{ij}). \tag{7}$$

Here, $squash(\cdot)$ refers to the squashing function. Unlike the definition in literature²⁴, this function has been augmented with non-linearity to more effectively capture the correlations between attention capsules on spatially invariant feature surfaces, as expressed in equation (8),

$$squash(s) = \frac{\|s\|^2}{0.4 + \|s\|^2} \cdot \frac{s}{\|s\|}. \tag{8}$$

The network is trained using the margin loss function L_j , which directs the length of the class capsules to approximate the probability of the correct classification.

$$L_j = T_j \max(0, m^+ - \|s_j\|)^2 + \lambda(1 - T_j) \max(0, \|s_j\| - m^-)^2. \tag{9}$$

In equation (9), m^+ , m^- and λ are hyperparameters, with values set as 0.9, 0.1 and 0.5, respectively. The target $T_j = 1$ is assigned a value of 1 when the predicted class is correct, and 0 otherwise.

-
- 1: Calculate the feature capsules: $s_i \leftarrow [conv^*(conv(X_{SP}))]^f$
 - 2: **for all** feature capsules in layer i **do**
 - 3: Calculate the capsule correlation prediction matrix q_{ij} by applying transformation matrix W_{ij}^q : $q_{ij} \leftarrow s_i * W_{ij}^q [l_f * l_c]$
 - 4: Calculate the key prediction matrix k_{ij} by applying transformation matrix W_{ij}^k : $k_{ij} \leftarrow s_i * W_{ij}^k [l_f * l_c]$
 - 5: Calculate the shape prediction matrix v_{ij} by applying transformation matrix W_{ij}^v : $v_{ij} \leftarrow s_i * W_{ij}^v [l_f * l_c]$
 - 6: **end for**
 - 7: Calculate the coupling coefficient matrix between the feature capsules s_i and class capsules s_j : $c_{ij} \leftarrow softmax\left(\frac{q_{ij} * (k_{ij})^T}{\sqrt{d_k}}\right)$
 - 8: **for all** class capsules in layer j **do**
 - 9: Calculate the class capsules s_j : $s_j \leftarrow squash(c_{ij} * v_{ij})$
 - 10: **end for**
 - 11: **return** s_j
-

Algorithm 1. SISA-CapsNet.

Interpretability description

SISA-CapsNet achieves interpretable clustering from feature capsules to class capsules through its spatially invariant capsule extraction module and capsule self-attention routing module. Its interpretability is demonstrated in the following ways: (a) Feature capsules are constrained to encode spatial features from fixed regions and are represented as vectors; (b) Class capsules are generated through the linear combination of feature capsules, facilitated by coupling coefficients derived from the self-attention mechanism. The coupling coefficient matrix offers an intuitive means of evaluating the activation level of each feature capsule during the routing process; (c) Spatially invariant capsules are positioned on feature surfaces that match the size of the input image, enabling pixel-level interpretability of the spatial feature encoding sources.

Building on this, a spatial feature activation weight mask is constructed using the feature capsule coupling coefficient matrix. This mask provides an intuitive visualization of the activation levels of feature capsules during

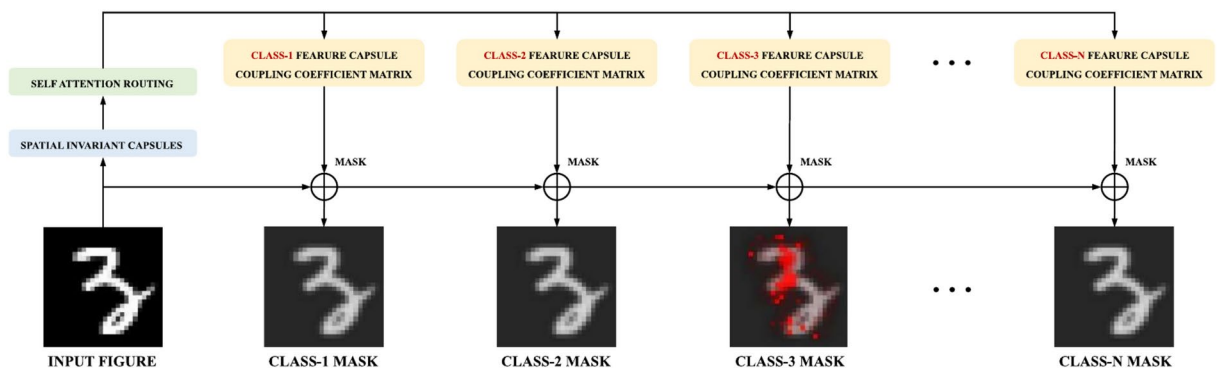


Fig. 4. Schematic representation of spatial feature activation weight mask application.

the classification task, as exemplified in Fig. 4 which demonstrates the application of this mask through a case study on the MNIST dataset.

For any given input image, the spatial feature activation weight mask is computed for different categories. Initially, the coupling coefficient matrix used to construct the mask is randomly initialized. As training progresses, the coupling coefficient matrix for incorrect categories fails to capture relevant features of the input image, leading to uniform activation across all feature capsules. In contrast, the coupling coefficient matrix for the correct category will strongly activate the feature capsules that align with the input image, while suppressing the activation of unrelated capsules. The coupling coefficient matrix is set as a red (R) channel mask and overlaid onto the input image using a Gaussian mixture. This approach allows for a visualized and interpretable representation of the classification process: In the mask corresponding to incorrect categories, all pixels exhibit the same average R channel value, resulting in a uniform gray mask. For the correct category, the regions with strong activations display high R channel values, while the non-activated regions exhibit low values. The stark contrast between these areas appears as red in the activated regions, providing an intuitive and human-readable rationale for the classification decision.

Experiments and results

Dataset

MNIST dataset

The MNIST dataset is a widely used benchmark for handwritten digit recognition, comprising 60,000 training samples and 10,000 test samples. Each sample is a 28×28 pixel grayscale image representing a digit label between 0 and 9. Due to its moderate scale, accurate annotations, and broad applicability, MNIST is commonly used to evaluate and compare the performance of various deep learning algorithms. It serves as an ideal dataset for validating the baseline performance of SISA-CapsNet model.

MSTAR dataset

The MSTAR (Mobile SAR Target Recognition) dataset is a widely used standard in synthetic aperture radar (SAR) image processing, primarily designed for target detection and recognition tasks, as shown in Fig. 5. It includes 10 types of military targets⁴⁸, such as the self-propelled howitzer (2S1), armored reconnaissance vehicle (BRDM2), and tank (T62). These targets are captured from various azimuth and elevation angles using X-band SAR imagery with a resolution of approximately 0.3 meters by 0.3 meters and an image size of 100×100 pixels. The MSTAR dataset is divided into training and testing sets, which were collected at 17° and 15° pitch angles, respectively. This provides a diverse range of target categories and imaging conditions, making it a reliable benchmark for SAR image recognition research. Due to factors like small sample size, inherent speckle noise, and geometric distortions, MSTAR presents challenges that make it ideal for testing the generalization and robustness of models like SISA-CapsNet, especially in the context of interpretable spatial feature-based classification.

USCS (Unmanned swarm combat situation) dataset

The Unmanned Swarm Combat Situation (USCS) image dataset is constructed based on battlefield situational dynamics in the confrontational processes of unmanned system swarms, as illustrated in Fig. 6. The dataset encapsulates situational information, including various combat unit types from both red and blue forces, intent trajectories, and operational interference. It is categorized into four situational classes: offensive, defensive, encirclement, and maneuver situations. Each image has a resolution of 77×77 pixels. By establishing a

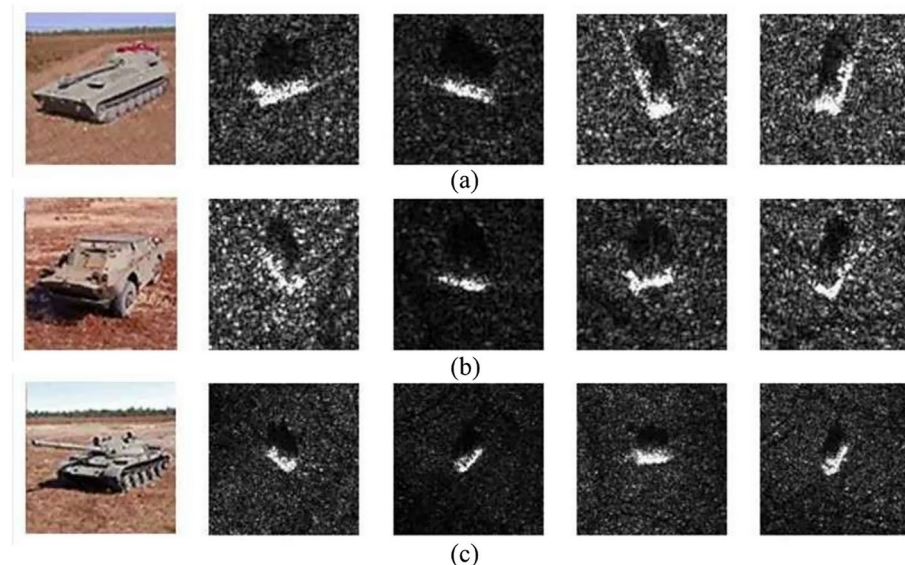


Fig. 5. The MSTAR image dataset (a) 2S1, (b) BRDM2, (c) T62.

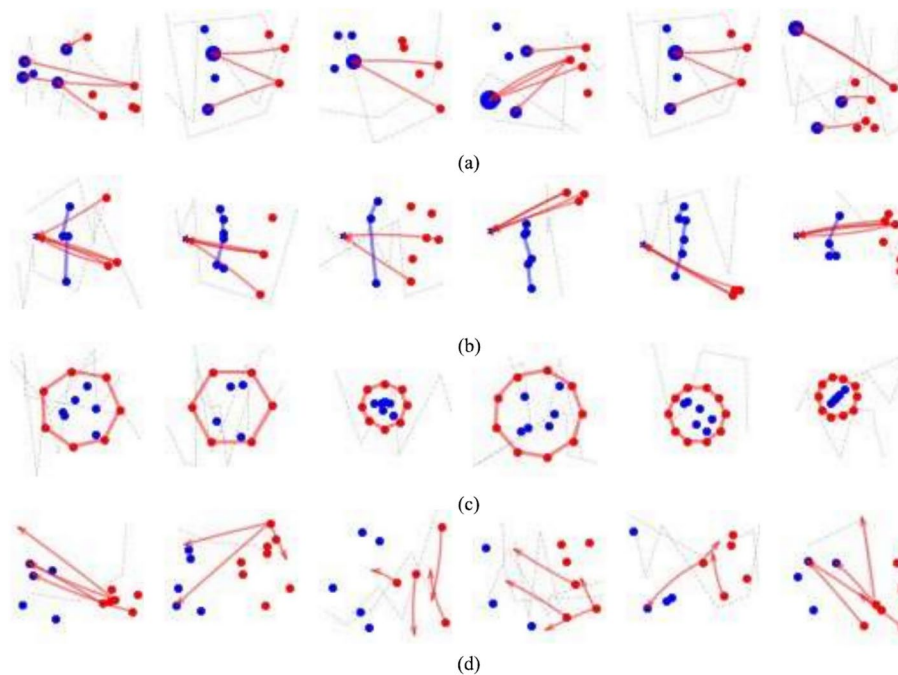


Fig. 6. The USCS image dataset (a) Offensive Situation, (b) Defensive Situation, (c) Encirclement Situation, (d) Maneuver Situation.

standardized visual representation framework, USCS serves as a reliable benchmark for unmanned swarm combat situation analysis and decision-making. Additionally, it facilitates the validation of SISA-CapsNet's explainability in situational image classification tasks.

- **Offensive Situation (Fig. 6a):** The red and blue combat units are randomly distributed in the right and left regions of the image, respectively. The assignment of offensive units and their corresponding targets follows a random selection mechanism. The red arrows indicate the direction of offensive, with blue units targeted by multiple red units represented by enlarged circular markers, signifying a heightened threat level.
- **Defensive Situation (Fig. 6b):** The blue base is denoted by a blue pentagram. Blue units are randomly positioned in the center of the image, forming a defensive line, while red units are scattered across the right region. The situational image clearly delineates the offensive vectors of red units targeting the blue base, as indicated by the offensive arrows.
- **Encirclement Situation (Fig. 6c):** Blue units form combat clusters at random locations within the image, while red units are positioned along the periphery, creating an encirclement. A polygonal encirclement is formed by connecting the red units with straight lines, visually illustrating the encirclement range and the tactical deployment strategy.
- **Maneuver Situation (Fig. 6d):** Both red and blue units are randomly positioned on opposite sides of the image. The red offensive strategy exhibits nondeterministic features, with some offensive directions targeting specific blue units, while others are directed toward random locations within the blue region. This strategy incorporates a range of tactical possibilities, including feigned offensives, flanking maneuvers, and firepower suppression.

Classification performance analysis

This section evaluates the classification accuracy of SISA-CapsNet using four benchmark datasets, with the deep learning networks ResNet-18 and AlexNet, Capsule Network (CapsNet)²⁴, as well as the self-attention-based capsule network Efficient-CapsNet²⁷ serving as comparative baseline models. The computational experiments were developed in PyTorch, with training executed on an Intel i5-9400F CPU paired with an NVIDIA GeForce RTX 2060 GPU.

MNIST performance

Figure 7 presents the training and testing accuracy of ResNet18, AlexNet, CapsNet, Efficient-CapsNet, and SISA-CapsNet on the MNIST dataset. After 100 epochs of training, the final training and testing accuracies were 98.72% and 98.73%, respectively. Compared to other deep learning-based and capsule-based networks, SISA-CapsNet exhibits a moderate gap in quantitative metrics such as classification accuracy. However, it maintains fundamental classification performance while demonstrating significant advantages in interpretability. In the next section, we will validate the interpretability improvement achieved by SISA-CapsNet, alongside its usability, through visualization methods.

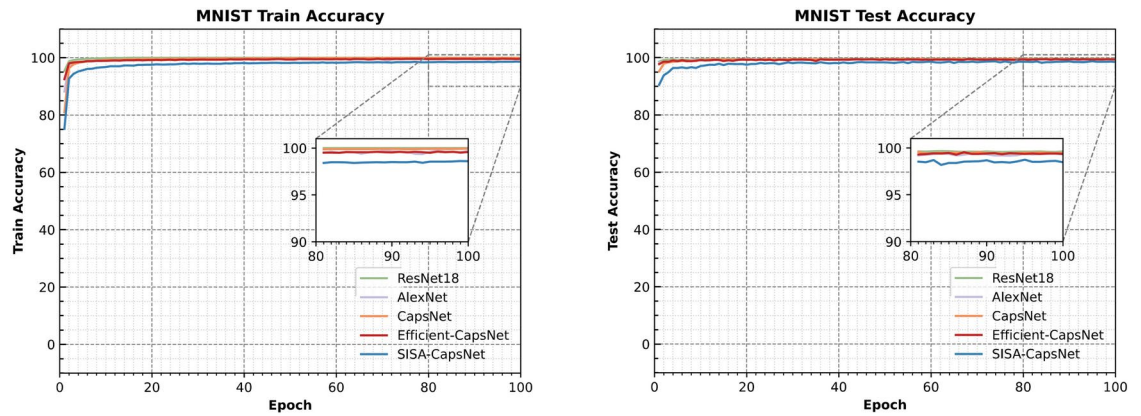


Fig. 7. Comparison of training and testing accuracy for ResNet-18, AlexNet, CapsNet, Efficient-CapsNet and SISA-CapsNet on the MNIST dataset.

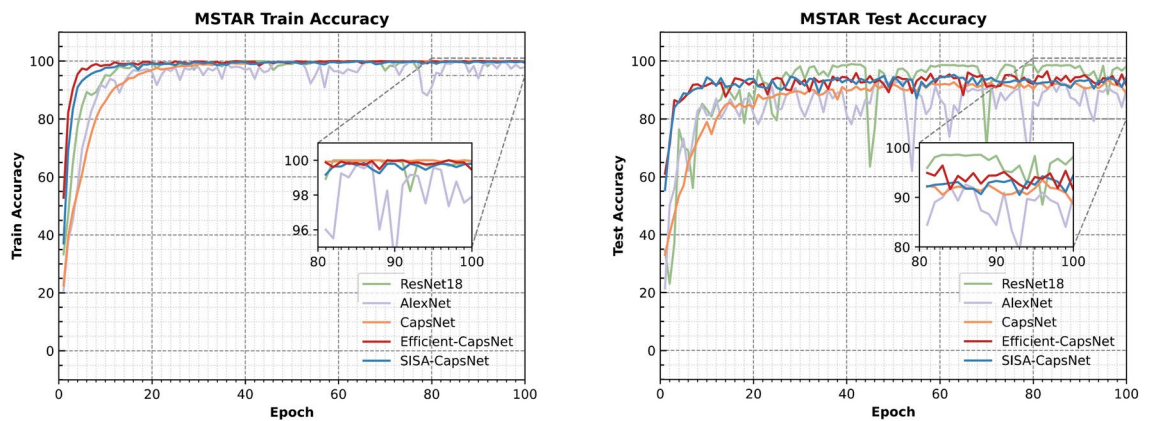


Fig. 8. Comparison of training and testing accuracy for ResNet-18, AlexNet, CapsNet, Efficient-CapsNet and SISA-CapsNet on the MSTAR dataset.

MSTAR performance

Figure 8 shows the training and testing accuracy of ResNet18, AlexNet, CapsNet, Efficient-CapsNet, and SISA-CapsNet on the MSTAR dataset. After 100 epochs of training, the final training and testing accuracies were 99.75% and 93.35%, respectively. On the training set, the networks performed similarly, with all networks successfully completing the MSTAR classification task. However, on the test set, ResNet and AlexNet exhibited significant fluctuations compared to the capsule-based networks, indicating that traditional deep learning architectures struggle with generalization and understanding when handling the high spatial feature demands and noise inherent in the MSTAR dataset. In contrast, SISA-CapsNet's feature encoding capabilities resulted in more stable performance, demonstrating stronger robustness on the small sample dataset.

USCS performance

Figure 9 illustrates the training and testing accuracy of ResNet18, AlexNet, CapsNet, Efficient-CapsNet, and SISA-CapsNet on the USCS dataset. After 100 epochs of training, the final training and testing accuracies were 99.99% and 99.93%, respectively. On the training set, the performance of various capsule-based networks was similar and superior to that of deep learning-based networks. However, on the test set, ResNet exhibited severe overfitting, rendering it incapable of performing situational image classification inference. Among the remaining networks, the proposed SISA-CapsNet demonstrated a more stable accuracy curve, highlighting its stronger generalization capability.

Additionally, we analyzed the parameter count of each network and measured the training time per batch across different datasets. These parameters, along with classification performance and interpretability metrics, are summarized in Table 1. As shown, the proposed SISA-CapsNet has a higher parameter count than other capsule networks. However, by utilizing parallel computation instead of iterative routing, it achieves reduced training time. Despite this improvement, SISA-CapsNet remains less efficient than EfficientNet, which leverages self-attention to enhance computational efficiency. This limitation becomes more pronounced on larger datasets but represents a necessary trade-off to achieve pixel-level interpretability.

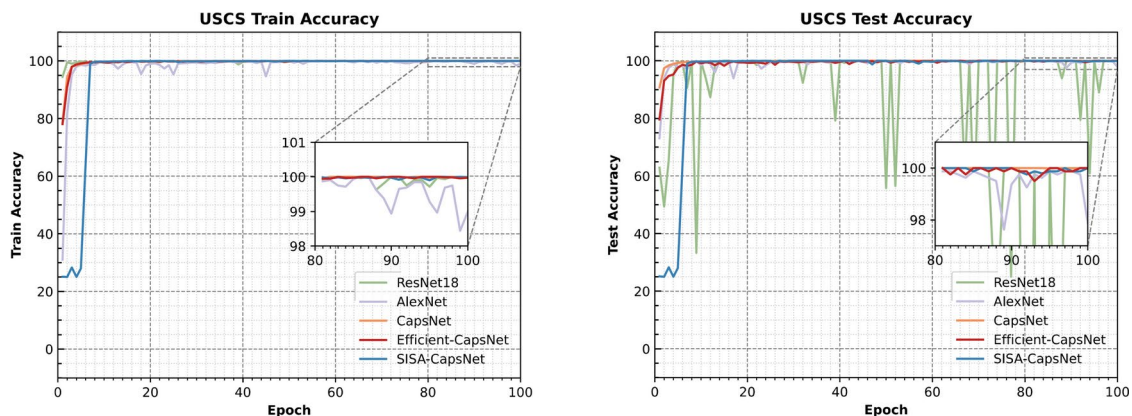


Fig. 9. Comparison of training and testing accuracy for ResNet-18, AlexNet, CapsNet, Efficient-CapsNet and SISA-CapsNet on the USCS dataset.

Model Name	Dataset Name	#Param. (M)	Batch Time (ms)	Training Accuracy (%)	Testing Accuracy (%)	GPU Friendliness	Spatial Interpretability
ResNet18	MNIST	11.18	15	100.00	99.59	Yes	No
	MSTAR	11.17	17	99.79	95.60	Yes	No
	USCS	11.17	17	99.98	97.81	Yes	No
AlexNet	MNIST	57.03	6	99.63	99.35	Yes	No
	MSTAR	57.03	6	98.21	88.70	Yes	No
	USCS	57.02	7	99.17	99.43	Yes	No
CapsNet	MNIST	6.80	24	99.92	99.48	No	No
	MSTAR	21.47	53	99.93	93.32	No	No
	USCS	5.90	22	99.98	99.91	No	No
Efficient-CapsNet	MNIST	1.48	12	99.57	99.39	Yes	No
	MSTAR	0.74	12	99.82	93.46	Yes	No
	USCS	1.18	11	99.98	99.95	Yes	No
SISA-CapsNet (Ours)	MNIST	8.85	11	98.72	98.73	Yes	Yes
	MSTAR	153.60	52	99.75	93.35	Yes	Yes
	USCS	32.74	18	99.99	99.93	Yes	Yes

Table 1. Comparison of algorithm performance across diverse datasets

Interpretability analysis

Leveraging the interpretable mask construction methodology detailed in Section [Methods](#), SISA-CapsNet is sequentially deployed on the MNIST, MSTAR, and USCS datasets, using feature capsule coupling coefficients to construct spatial feature activation weight masks. Figure 10 presents the inference results for the MNIST dataset. In the class mask, only the correctly classified category is prominently activated and highlighted in red, while the masks for other categories remain uniform. For example, in the case of the digits 3 and 4, SISA-CapsNet effectively identifies the distinctive spatial features of these digits from various angles and forms, utilizing these features as key classification criteria. This process closely aligns with human interpretation, offering a clear and interpretable reasoning for the classification. Specifically, the curvature in the upper part of the digit 3 serves as a unique feature that distinguishes it from other categories, while the distinct starting and crossing points in the digit 4 represent unique features specific to that category.

Figure 11 displays the inference results on the MSTAR dataset. As with the previous example, the significant activation areas are visible in the class mask. It is apparent that SISA-CapsNet effectively differentiates between various vehicles based on the shadow and shape characteristics inherent in the SAR images.

Figure 12 illustrates the inference results for the USCS dataset. From the figure, it is evident that SISA-CapsNet effectively identifies the key distinguishing features for different categories: (a) In the Offensive Situation, SISA-CapsNet extracts the critical feature of the intent curve, which only appears between combat units, differentiating it from the Maneuver Situation, where the combat units themselves are treated as the primary features in the Offensive Situation; (b) In the Defensive Situation, SISA-CapsNet recognizes the defensive intent curve as the key feature, while the offensive intent curve is regarded as irrelevant, showcasing its ability to distinguish between primary and secondary spatial features; (c) In the Encirclement Situation, SISA-CapsNet identifies both the combat units and the encirclement intent curve as the significant features for this category of images; (d) In

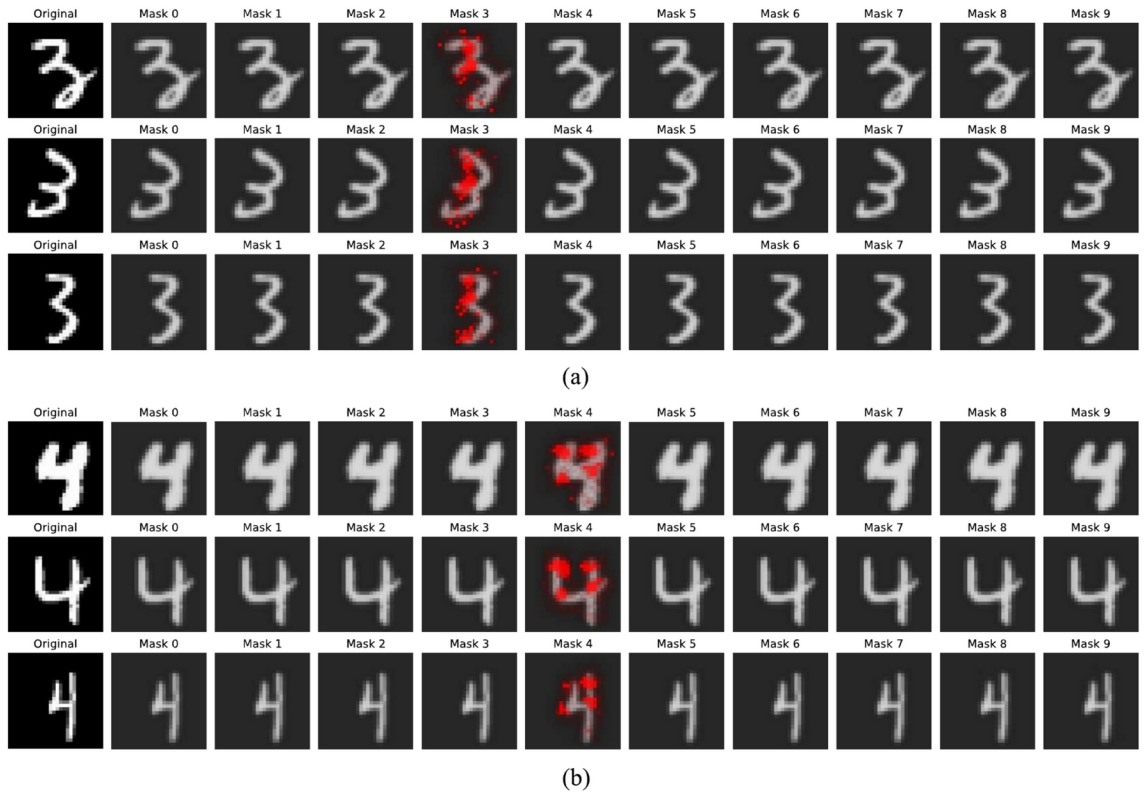


Fig. 10. Interpretability of SISA-CapsNet on the MNIST dataset visualized using spatial feature activation weight masks (a) Handwritten Digit 3, (b) Handwritten Digit 4.

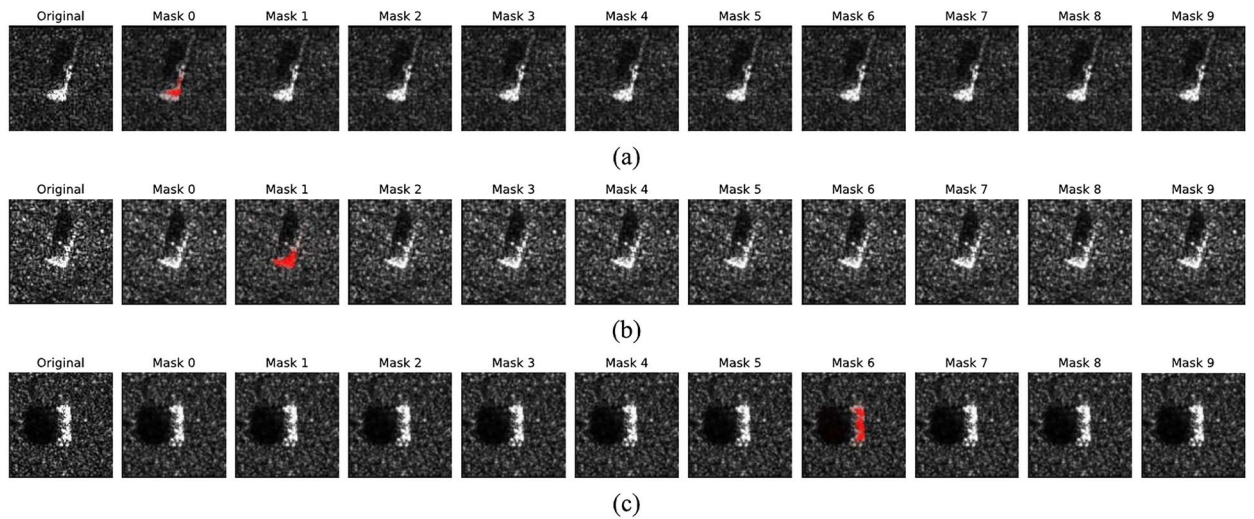


Fig. 11. Interpretability of SISA-CapsNet on the MSTAR dataset visualized using spatial feature activation weight masks (a) 2S1, (b) BRDM2, (c) T62.

the Maneuver Situation, SISA-CapsNet successfully highlights the non-directional intent curve as the significant feature of this category.

In summary, across multiple datasets, SISA-CapsNet effectively extracts spatial features that carry interpretable semantic information, aligning well with human understanding. This provides a safer and more reliable method for interpretable situational classification, offering clear decision rationale that is accessible to human users.

To further validate the interpretability of SISA-CapsNet in situational image classification, a comparison experiment was conducted using the widely recognized explainable image classification method, Class Activation Mapping (CAM)¹⁷, on the USCS dataset. CAM generates a heatmap related to a specific class by

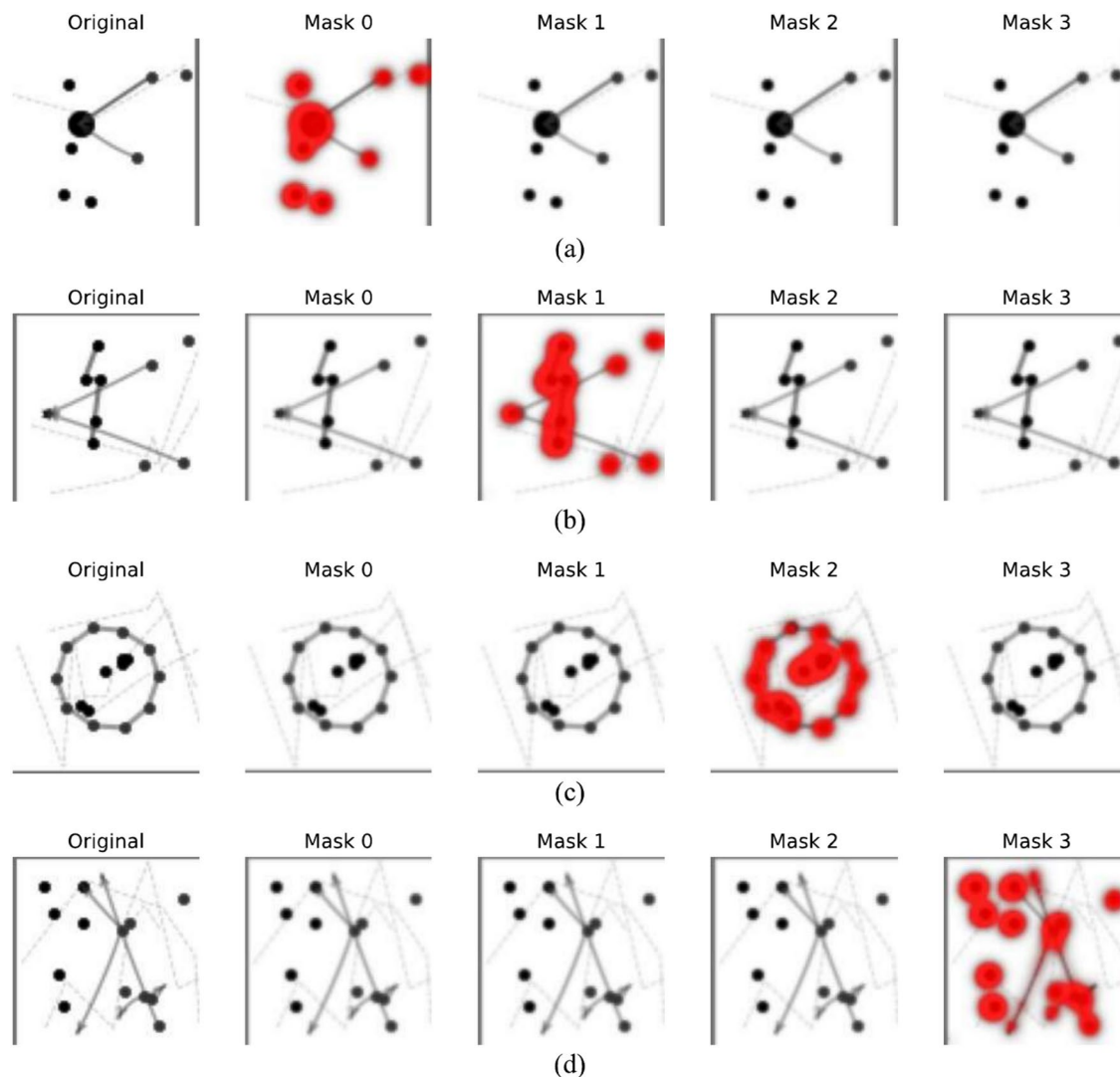


Fig. 12. Interpretability of SISA-CapsNet on the UCS dataset visualized using spatial feature activation weight masks (a) Offensive Situation, (b) Defensive Situation, (c) Encirclement Situation, (d) Maneuver Situation.

extracting and combining the weighted features from the final convolutional feature surfaces of a convolutional neural network. This heatmap visually highlights the key areas in the input image that influence the classification decision. However, CAM has several limitations: (a) **Global Average Pooling Issue.** CAM relies on the global average pooling operation, which often causes feature loss or mixing. This makes it difficult to accurately localize fine-grained spatial features. (b) **Network Structure Dependence.** CAM is limited by its reliance on specific network architectures, such as CNNs with global average pooling layers. To achieve accurate interpretability, the baseline network must perform well on the situational image classification task.

Figure 13 compares the interpretability of CAM's activation heatmap with SISA-CapsNet's spatial feature activation weight mask on the UCS dataset. From this comparison, we observe that CAM's activation areas do not align well with human understanding. Specifically, we note:

- **Interference Activation:** CAM incorrectly activates irrelevant background elements, such as lines in the Offensive and Defensive Situations.
- **Inability to Recognize Typical Features:** CAM fails to effectively identify key features of certain categories, such as the encirclement curve in the encirclement scenario or the movement intention curve in the mobility scenario.

In contrast, SISA-CapsNet aligns much better with human interpretation, providing clearer and more accurate explanations for the classification decisions.

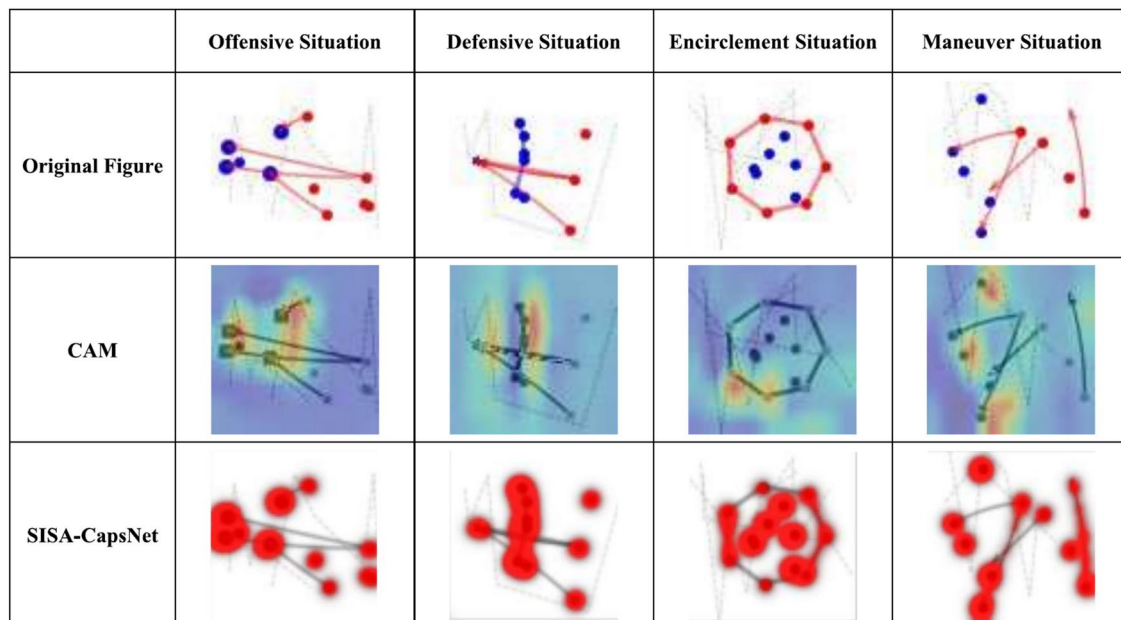


Fig. 13. Comparative interpretability of CAM and SISA-CapsNet on the USCS dataset.

	SISA-CapsNet	SI-CapsNet	SA-CapsNet
Training Accuracy (%)	98.72	98.04	97.39
Testing Accuracy (%)	98.73	98.15	96.85

Table 2. Ablation study results of SISA-CapsNet on classification performance.

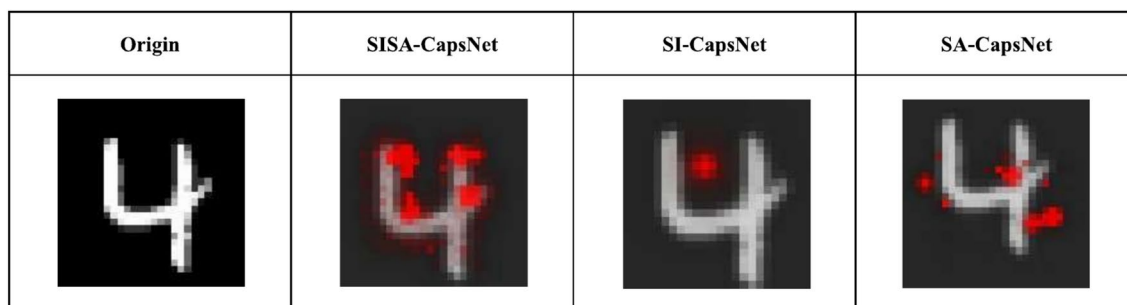


Fig. 14. Ablation study results of SISA-CapsNet on interpretability.

Ablation study

This section validates the contributions of the spatially invariant capsule extraction module (SI-Generator) and the capsule self-attention routing module (SA-Routing) in SISA-CapsNet to overall network classification performance and interpretability through ablation experiments. Specifically, we construct two ablation networks, SI-CapsNet and SA-CapsNet, to examine their effects on classification and interpretability performance. SI-CapsNet retains SI-Generator while replacing SA-Routing with the dynamic routing mechanism of the original capsule network. In contrast, SA-CapsNet preserves SA-Routing while substituting SI-Generator with traditional convolution to generate feature capsules. Additionally, to assess interpretability, the feature planes in SA-CapsNet are set to match the input image dimensions. Table 2 presents the results of the ablation experiments on the MNIST dataset, demonstrating that removing any individual module leads to a certain degree of degradation in the classification performance of SISA-CapsNet.

Furthermore, we focus primarily on the impact of ablation on the interpretability of SISA-CapsNet, as this is the key objective in designing the network architecture. Figure 14 displays the results of the ablation experiments on interpretability using the MNIST dataset, revealing that only SISA-CapsNet can effectively extract key interpretable semantic features from handwritten digits. In contrast, networks with any ablated module exhibit

unclear and difficult-to-interpret classification criteria. These findings indicate that both SI-CapsNet and SA-CapsNet play an indispensable role in achieving the interpretability of SISA-CapsNet.

Conclusion

This paper proposes a novel Spatial-Invariant Self-Attention Capsule Network (SISA-CapsNet) to achieve explainable situational image classification. The network encodes interpretable features with clear semantic information using spatial-invariant capsules, and performs classification by linearly combining features via capsule self-attention routing, enabling explainable classification of situational images. Experiments were conducted on several datasets: the handwritten digit dataset (MNIST), the SAR image military target dataset (MSTAR), and the situational image dataset (USCS). The results show that SISA-CapsNet effectively identifies significant spatial features in the USCS dataset, which contains rich semantic information and prominent spatial features, and successfully performs situational classification. Compared to CapsNet and traditional deep neural networks, SISA-CapsNet demonstrates more stable performance. Moreover, it outperforms CAM in terms of clearer and more intuitive interpretability, providing a reliable solution for situational image classification.

However, our method employs a substantial parameter volume and entails significant computational complexity to achieve pixel-level interpretability. This limitation becomes increasingly pronounced as image dimensions expand, revealing an inherent trade-off between fine-grained interpretability and computational efficiency. Moving forward, we aim to further investigate potential solutions to this challenge. Prospective approaches may include introducing sparsity constraints or implementing hierarchical architectures to selectively activate key feature capsules. Such optimizations would maintain the network's interpretability while enhancing computational efficiency, thereby improving applicability in higher-resolution imaging scenarios.

Data availability

The MNIST datasets used and/or analysed during the current study are available from the MNIST repository. (<https://github.com/Rujiyuan/MNIST-Dataset>) The MSTAR datasets used and/or analysed during the current study are available from the MSTAR repository. (<https://www.sdms.afrl.af.mil/index.php?collection=mstar>) The USCS datasets generated and/or analysed during the current study are available in the Rujiyuan repository. (<https://github.com/Rujiyuan/USCS-Dataset>)

Code availability

The Code used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 29 January 2025; Accepted: 1 April 2025

Published online: 15 April 2025

References

- Chen, R., Li, H., Yan, G., Peng, H. & Zhang, Q. Joint situational assessment-hierarchical decision-making framework for maneuver intent decisions. *Advanced Intelligent Systems*. 2300574 (2024).
- Munir, A., Aved, A. & Blasch, E. Situational awareness: techniques, challenges, and prospects. *AI* **3**, 55–77 (2022).
- Li, Q. & Wang, X. Image classification based on sift and svm. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. 762–765 (IEEE, 2018).
- Tsourounis, D., Kastaniotis, D., Theoharatos, C., Kazantzidis, A. & Economou, G. Sift-cnn: when convolutional neural networks meet dense sift descriptors for image and sequence classification. *Journal of Imaging*. **8**, 256 (2022).
- Zhao, Z.-Q., Zheng, P., Xu, S.-T. & Wu, X. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*. **30**, 3212–3232 (2019).
- Zhang, Y., Zhang, T., Wang, S. & Yu, P. An efficient perceptual video compression scheme based on deep learning-assisted video saliency and just noticeable distortion. *Engineering Applications of Artificial Intelligence*. **141**, 109806 (2025).
- Zhang, Y., Wang, S., Zhang, Y. & Yu, P. Asymmetric light-aware progressive decoding network for rgb-thermal salient object detection. *Journal of Electronic Imaging*. **34**, 013005–013005 (2025).
- Hasan, M. A., Bhargav, T., Sandeep, V., Reddy, V. S. & Ajay, R. Image classification using convolutional neural networks. *International Journal of Mechanical Engineering Research and Technology*. **16**, 173–181 (2024).
- Peng, H., Zhang, Y., Yang, S. & Song, B. Battlefield image situational awareness application based on deep learning. *IEEE intelligent systems*. **35**, 36–43 (2019).
- Zhang, Y., Zhen, J., Liu, T., Yang, Y. & Cheng, Y. Adaptive differentiation siamese fusion network for remote sensing change detection. *IEEE Geoscience and Remote Sensing Letters*. (2024).
- Zhang, Y., Wu, C., Guo, W., Zhang, T. & Li, W. Cfanet: Efficient detection of uav image based on cross-layer feature aggregation. *IEEE Transactions on Geoscience and Remote Sensing*. **61**, 1–11 (2023).
- Zhang, Y., Wu, C., Zhang, T. & Zheng, Y. Full-scale feature aggregation and grouping feature reconstruction based uav image target detection. *IEEE Transactions on Geoscience and Remote Sensing*. (2024).
- Wani, N. A., Kumar, R. & Bedi, J. Deepexplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*. **243**, 107879 (2024).
- Van der Velden, B. H., Kuijff, H. J., Gilhuijs, K. G. & Viergever, M. A. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*. **79**, 102470 (2022).
- Haufe, S. et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*. **87**, 96–110 (2014).
- Kovalerchuk, B., Dunn, A., Worland, A. & Wagle, S. Interactive decision tree creation and enhancement with complete visualization for explainable modeling. In *Artificial Intelligence and Visualization: Advancing Visual Knowledge Discovery*, 3–40 (Springer, 2024).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929 (2016).
- Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626 (2017).
- Moshe, O., Fidel, G., Bitton, R. & Shabtai, A. Improving interpretability via regularization of neural activation sensitivity. *Machine Learning*. **113**, 6165–6196 (2024).

20. Ribeiro, M. T., Singh, S. & Guestrin, C. “ why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144 (2016).
21. Lundberg, S. A unified approach to interpreting model predictions. *arXiv preprint*, [arXiv:1705.07874](https://arxiv.org/abs/1705.07874) (2017).
22. Sun, T., Chen, H., Hu, G. & Zhao, C. Explainability-based knowledge distillation. *Pattern Recognition*. **159**, 111095 (2025).
23. Hinton, G. E., Krizhevsky, A. & Wang, S. D. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14–17, 2011, Proceedings, Part I 21*, 44–51 (Springer, 2011).
24. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. *Advances in neural information processing systems*. **30** (2017).
25. Shi, R. & Niu, L. A brief survey on capsule network. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 682–686 (IEEE, 2020).
26. Rajasegaran, J. et al. Deepcaps: Going deeper with capsule networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10725–10733 (2019).
27. Mazzia, V., Salvetti, F. & Chiaberge, M. Efficient-capsnet: Capsule network with self-attention routing. *Scientific reports*. **11**, 14634 (2021).
28. Gu, J. & Tresp, V. Improving the robustness of capsule networks to image affine transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7285–7293 (2020).
29. Liu, S. et al. Eeg emotion recognition based on the attention mechanism and pre-trained convolution capsule network. *Knowledge-Based Systems*. **265**, 110372 (2023).
30. Huang, W. & Zhou, F. Da-capsnet: dual attention mechanism capsule network. *Scientific Reports*. **10**, 1–13 (2020).
31. Jung, D., Lee, J., Yi, J. & Yoon, S. icaps: An interpretable classifier via disentangled capsule networks. In *European Conference on Computer Vision*, 314–330 (Springer, 2020).
32. Xi, E., Bing, S. & Jin, Y. Capsule network performance on complex data. *arXiv preprint*. [arXiv:1712.03480](https://arxiv.org/abs/1712.03480) (2017).
33. Afshar, P., Plataniotis, K. N. & Mohammadi, A. Capsule networks’ interpretability for brain tumor classification via radiomics analyses. In *2019 IEEE International Conference on Image Processing (ICIP)*, 3816–3820 (IEEE, 2019).
34. Sezavar, A., Atta, R. & Ghanbari, M. Dcapsnet: Deep capsule network for human activity and gait recognition with smartphone sensors. *Pattern Recognition*. **147**, 110054 (2024).
35. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*. (2017).
36. Zhang, Y., Liu, Y., Kang, W. & Tao, R. Vss-net: Visual semantic self-mining network for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*. **34**, 2775–2788 (2023).
37. Zhang, Y., Zhang, T., Wu, C. & Tao, R. Multi-scale spatiotemporal feature fusion network for video saliency prediction. *IEEE Transactions on Multimedia*. **26**, 4183–4193 (2023).
38. Liu, H.-I. & Chen, W.-L. X-transformer: a machine translation model enhanced by the self-attention mechanism. *Applied Sciences*. **12**, 4502 (2022).
39. Parmar, N. et al. Image transformer. In *International conference on machine learning*. 4055–4064 (PMLR, 2018).
40. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020).
41. Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M. & Hu, S.-M. Visual attention network. *Computational visual media*. **9**, 733–752 (2023).
42. Cheng, G., Lai, P., Gao, D. & Han, J. Class attention network for image recognition. *Science China Information Sciences*. **66**, 132105 (2023).
43. Wu, W., Liu, S., Xia, Y. & Zhang, Y. Dual residual attention network for image denoising. *Pattern Recognition*. **149**, 110291 (2024).
44. Hoogi, A., Wilcox, B., Gupta, Y. & Rubin, D. L. Self-attention capsule networks for object classification. *arXiv preprint* [arXiv:1904.12483](https://arxiv.org/abs/1904.12483) (2019).
45. Yu, X. et al. Research on a capsule network text classification method with a self-attention mechanism. *Symmetry* **16**, 517 (2024).
46. Ke, S., Ma, C., Li, W., Lv, J. & Zou, L. Multi-region and multi-band electroencephalogram emotion recognition based on self-attention and capsule network. *Applied Sciences*. **14**, 702 (2024).
47. Hinton, G. E., Sabour, S. & Frosst, N. Matrix capsules with em routing. In *International conference on learning representations*. (2018).
48. Gu, Y., Tao, J., Feng, L. & Wang, H. Using vgg16 to military target classification on mstar dataset. In *2021 2nd China International SAR Symposium (CISS)*. 1–3 (IEEE, 2021).

Author contributions

P.Z.L, B.W., Q.F. and Z.C. conceived of the study, P.Z.L and J.Y.R. conducted the experiments, J.Y.R. and P.Z.L participated in the interpretation of the results and drafting/revising the manuscript, B.W., Q.F. and Z.C. helped to revise the manuscript critically. All the authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025