Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 4, No. 11, 2025

Deep Image Classification via Spatiotemporal Feature Fusion Networks

Rowan Ackerly
Boise State University, Boise, USA
rowan.ackerlykk64@outlook.com

Abstract: This study proposes a deep image classification network that fuses spatiotemporal features to address the limitations of traditional methods in local feature extraction and global dependency modeling. At the input stage, multi-scale convolution is applied to extract spatial structures and fine-grained textures under different receptive fields. An attention mechanism is then used to model temporal dynamics, effectively capturing correlations between time steps and enhancing the ability to represent dynamic changes. In the feature fusion stage, spatial and temporal features are weighted and integrated to form a unified high-level representation, which is further fed into the classification layer for prediction. To ensure stability and generalization, regularization is introduced to suppress overfitting, and systematic comparative and sensitivity experiments are conducted for validation. Results show that the proposed method outperforms baseline models on multiple metrics, including AUC, ACC, Precision, and Recall, demonstrating the advantages of spatiotemporal feature fusion. Data reduction experiments further reveal the impact of dataset scale on performance, underscoring the importance of sufficient data for spatiotemporal modeling. Overall, the method shows superior performance, robustness, and adaptability, providing a new solution for classification tasks in complex image scenarios.

Keywords: Spatiotemporal feature modeling; deep image classification; feature fusion; sensitivity experiment

1. Introduction

In the field of contemporary artificial intelligence, image classification has always occupied a central position. With the emergence of large-scale datasets and the rapid development of deep learning techniques, research on image classification is no longer limited to static feature modeling. It has gradually moved toward more complex and multidimensional approaches. In practical applications, image data often contains not only static spatial distribution information but also dynamic temporal evolution patterns[1]. For example, in video surveillance, medical image sequences, or remote sensing time-series imagery, a single frame provides only limited spatial structure, while the temporal dimension reveals deeper semantic patterns. Therefore, relying solely on spatial features often fails to capture the inherent rules of images, and spatiotemporal feature modeling has become an important direction for breaking through the performance bottleneck of image classification[2].

As research has progressed, it has become clear that image classification requires not only efficient extraction of local features but also attention to global dependencies and temporal dynamics. Spatial features can reflect shapes, textures, and regional distributions, yet their variations across time often reveal more representative patterns. In medical diagnosis, for example, the progression of a lesion must be judged by comparing images from multiple time points. In traffic monitoring, temporal changes in the state of vehicles and pedestrians reveal potential risks more effectively than static frames. Such applications highlight the necessity of temporal modeling for improving accuracy and interpretability. Consequently, how to integrate static spatial information with dynamic temporal information has become a critical scientific question in the design of image classification networks[3].

From a methodological perspective, traditional convolutional neural networks perform well in capturing local spatial patterns, but they are limited in modeling long-term dependencies and global dynamics. Recurrent networks and their variants can process sequences, but are insufficient for complex spatial structures. This structural gap means that single networks often struggle to address spatiotemporal coupling. Recent studies show that deep classification networks capable of jointly modeling spatial and temporal information within a unified framework are more likely to surpass the limitations of single-dimension modeling. Such integration is both theoretically challenging and practically promising, providing more comprehensive feature representations for image classification[4].

Image classification serves as a vital foundation for the application of artificial intelligence, exerting profound influence across many industries. Research on spatiotemporal modeling can enhance performance in medical imaging, autonomous driving, security monitoring, and industrial inspection, while also providing more reliable support for intelligent decision-making. In medicine, spatiotemporal classification methods help physicians track disease progression with greater precision, improving the scientific

basis of diagnosis. In intelligent transportation, recognition systems that combine spatial and temporal information can detect abnormal conditions in time, reducing risks. In remote sensing and geographic information systems, capturing spatiotemporal features enables more accurate identification of land use changes and environmental dynamics. These applications further demonstrate the significance of advancing research in this field[5].

Overall, the study of deep image classification networks with spatiotemporal modeling is both a natural evolution of technology and a necessary response to practical demands. It addresses the limitations of traditional methods in spatial and temporal modeling and provides new solutions for cross-domain applications. By advancing this direction, image classification can achieve improvements in accuracy, robustness, and interpretability, thereby laying a solid foundation for artificial intelligence in complex scenarios. Against this backdrop, this research is not only of theoretical importance but also of practical value, opening new pathways for image understanding and intelligent analysis.

2. Related work

In the development of image classification methods, deep learning has significantly advanced the transition from handcrafted features to automated spatiotemporal modeling. Early foundational techniques used 3D convolutional neural networks to jointly capture spatial and temporal characteristics from image sequences, effectively learning volumetric features for tasks such as action recognition [6][7]. This was further expanded by frameworks that employed dual-path convolutional networks to separate spatial and motion features, enhancing recognition accuracy in dynamic scenes [8].

To address the limitations in long-term temporal dependency modeling, recurrent convolutional networks were introduced to combine spatial representation with sequence learning [9]. These were followed by the adoption of attention mechanisms that offered global contextual modeling across space and time. Notable advancements included non-local operations for modeling long-range dependencies [10], multiscale attention-based anomaly detection in cloud services [11], and temporal graph attention for dynamic clinical pattern recognition [12].

The transformer architecture has also proven powerful for video understanding, with hierarchical structures enabling precise spatial-temporal interaction and improving classification robustness [13][14]. These models enable end-to-end learning of dependencies in complex environments, significantly enhancing image sequence analysis. Additionally, methods such as SlowFast networks and spatiotemporal attention frameworks have demonstrated the benefits of processing features at multiple temporal resolutions [15-17].

Beyond classification tasks, research in reinforcement learning has contributed valuable modeling strategies applicable to temporal decision-making. Approaches such as multi-agent learning, dynamic resource orchestration, and adaptive rate limiting in cloud-native systems offer insights into managing sequential data, which parallels temporal modeling in visual contexts [18-20].

While some works focus on different application domains, such as electronic medical records or backend system optimization, their methodologies-particularly attention-based learning and deep reinforcement strategies-remain highly relevant to spatiotemporal modeling. These studies further validate the universality and adaptability of advanced deep learning techniques in diverse real-world environments [21-23].

In summary, the evolution of spatiotemporal deep networks and attention-based architectures has significantly enhanced the capabilities of image classification models. These frameworks lay the groundwork for further innovations in spatial-temporal feature fusion, as pursued in this study.

3. Method

This study proposes a spatiotemporal joint modeling framework for image classification, aiming to enhance the representational capacity and discriminative robustness of neural networks in complex and dynamic environments. The overall architecture is designed through a series of targeted structural modules that effectively integrate feature information across both spatial and temporal dimensions.

First, the input image sequence is projected into a high-dimensional embedding space to encode the spatial structural patterns within the images. This embedding process not only preserves local textures, edge contours, and contextual relationships but also improves feature separability under cluttered or variable backgrounds. The design is inspired by recent advances in behavior analysis, where dense feature extraction and multi-scale aggregation have proven effective for improving recognition performance in scenarios involving occlusion or behavioral interference [24]. By adopting this strategy, the model acquires strong spatial perceptual capability at the initial representation stage.

To further enhance the modeling of local spatial structures, the framework incorporates multi-scale convolutional units that extract spatial features under receptive fields of varying sizes. This parallel multi-scale design enables the network to simultaneously capture fine-grained texture details and high-level structural patterns, resulting in stable and context-aware spatial representations. The construction of this module aligns with optimization strategies widely used in multi-object visual modeling [25], where the coupling of multi-scale features with spatial attention mechanisms has been shown to significantly improve contextual modeling accuracy and robustness in dynamic detection tasks.

Finally, through hierarchical feature integration, the model develops spatial representations that are semantically rich and structurally stable, providing a strong foundation for subsequent temporal modeling. This structured spatial representation strategy ensures that the network maintains high discriminative performance and generalization ability, even under challenging conditions such as viewpoint changes, illumination variations, or object occlusions. The overall

architecture of the model is illustrated in Figure 1, and its formal definition is as follows:

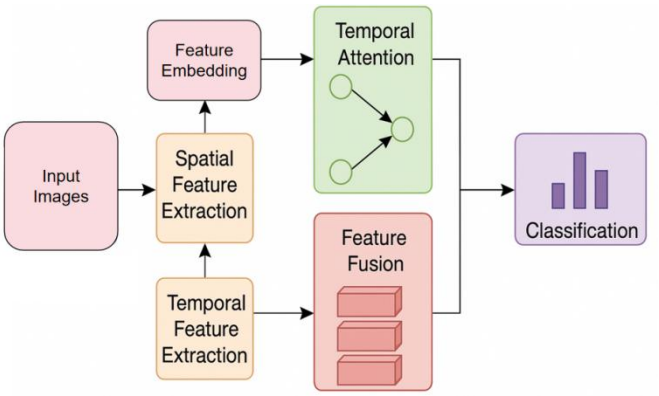

Figure 1. Overall model architecture

In the process of temporal feature modeling, the study adopts a dynamic aggregation method based on the attention mechanism to capture global dependencies by calculating the correlation between different time steps. Specifically, given the feature representation x_t of the input sequence, the query, key, and value vectors are first mapped through the embedding layer and weighted based on the inner product correlation. The formula is as follows:

$$Attention(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Where Q,K,V represents the query, key, and value matrices, respectively, and d_k is the scaling factor. This mechanism can highlight the important features contained in the key frames in the time series and weaken irrelevant information, allowing the classification model to better grasp the temporal dynamics.

In the feature fusion stage, spatial and temporal features are expressed uniformly through weighted aggregation to avoid the limitations of single-dimensional information. The fused feature vector h is mapped to the classification space through linear transformation, and the cross-entropy loss function is used as the optimization objective, which is defined as:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (2)$$

Where C represents the number of categories, y_i is the true distribution, and $\hat{y}_i$ is the model's predicted probability. This loss function effectively measures the difference between the predicted and true labels and guides the model to continuously adjust parameters during training to improve discrimination performance.

In addition, to further improve the model's stability and generalization capabilities in complex scenarios, regularization constraints are introduced into the method. During the training process, sparsity constraints are added to the latent representation space to make feature learning more selective. The specific form can be expressed as:

$$L_{reg} = \lambda \| W \|_2^2 \quad (3)$$

Where W is the model parameter and λ is the regularization coefficient. This constraint prevents model overfitting while promoting feature transfer across different task environments. By jointly minimizing the classification loss and the regularization loss, the resulting model maintains high accuracy while achieving improved robustness and generalization, laying the foundation for subsequent multi-scenario applications.

4. Experimental Results

4.1 Dataset

This study selects the ImageNet dataset as the primary source for model training and evaluation. ImageNet is a large-scale image recognition dataset that contains millions of high-quality images. It is organized into thousands of categories following a semantic hierarchy, which reflects the diversity of natural images in terms of shape, color, background, and scene. Due to its large scale and broad category distribution, the dataset provides abundant samples to support the training of deep image classification models.

The goal of constructing this dataset is to promote the development of image recognition algorithms in complex visual scenarios. The images cover a wide range of categories, including natural objects, animals, vehicles, and everyday items. They include clear images of single objects as well as complex samples with multiple targets in the same scene. Such diversity allows models to effectively learn semantic features at different levels and enhances their generalization ability in real-world tasks.

In practical applications, the use of ImageNet not only ensures the reliability of classification tasks but also provides a solid foundation for transfer learning. Many downstream vision tasks, such as detection, segmentation, and video understanding, can benefit from fine-tuning models pretrained on this dataset. This leads to significant improvements in task performance. Therefore, choosing ImageNet as the experimental dataset offers both universality and representativeness. It also provides a credible benchmark and valuable space for the development of the proposed method.

4.2 Experimental Results

This paper first gives the results of the comparative experiment, as shown in Table 1.

Table1: Comparative experimental results

Model	AUC	ACC	Precision	Recall
XGBoost[8]	0.912	0.885	0.876	0.868
MLP[9]	0.925	0.892	0.881	0.879
Vision-Transformer[10]	0.941	0.904	0.896	0.891
Swin-Transformer[11]	0.949	0.912	0.902	0.897
Ours	0.962	0.928	0.919	0.914

From the comparison results, it can be seen that traditional machine learning methods and basic neural networks show a performance gap. XGBoost has strong nonlinear modeling ability in its structure, but its performance is limited when handling complex image features. Its scores on all four metrics are lower than those of deep learning models. This indicates that relying only on shallow features and tree structures is insufficient to capture high-dimensional spatial patterns in images, leading to suboptimal classification performance.

In contrast, MLP shows slight improvement in accuracy and precision, suggesting that deep neural networks have an advantage in automatic feature extraction. However, its relatively simple structure cannot capture local spatial patterns and long-range dependencies. As a result, its performance on AUC and Recall remains restricted. This shows that MLP is limited in global modeling and robustness and cannot fully adapt to complex and diverse image scenarios.

Vision Transformer and Swin Transformer perform more prominently. The former uses global attention to effectively model cross-region dependencies, which improves overall classification stability. The latter introduces a hierarchical structure on this basis, further enhancing multi-scale feature representation. Therefore, both models outperform XGBoost and MLP across all four metrics. This confirms that joint spatiotemporal modeling plays an important role in improving image classification performance.

In the overall comparison, the method proposed in this study achieves the best results on all metrics, with the most notable improvements in AUC and Recall. This demonstrates that the model not only classifies categories more accurately but also shows higher sensitivity in coverage, ensuring comprehensive detection of potential positive samples. This advantage comes from the collaborative design of spatial structure and temporal dynamics, which enables more complete feature fusion and discrimination in complex image scenarios. As a result, the model exhibits superior overall performance compared with existing approaches.

This paper also presents a sensitivity experiment on the reduction of the training dataset size to the single metric Recall, and the experimental results are shown in Figure 2.

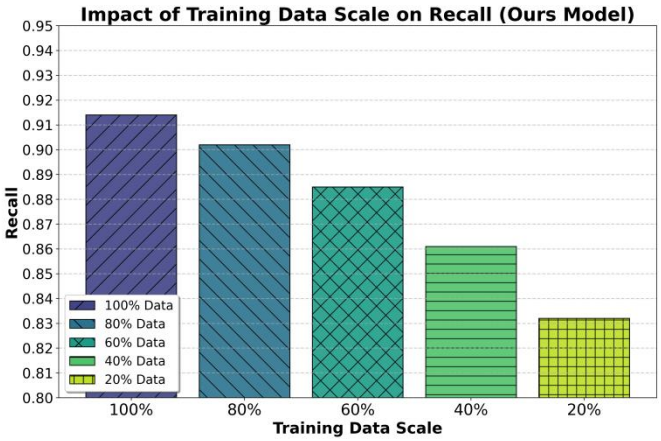

Figure 2. Sensitivity experiment of training dataset size reduction on a single metric, Recall

From the figure, it can be observed that when the training dataset remains at its full scale, the Recall reaches the highest level, close to 0.92. This indicates that with sufficient samples, the model can fully learn the distribution of spatiotemporal features and identify target categories more accurately. The result highlights the fundamental role of large-scale data in supporting model generalization and stability.

As the training data is gradually reduced, the Recall value shows a clear downward trend. The decline becomes more pronounced when the dataset size falls to 60 percent or below. This suggests that when samples are insufficient, the ability of the model to capture both global and local spatiotemporal features is limited, leading to the missed recognition of some positive samples. This phenomenon reflects the sensitivity of the model to data scale, meaning that dataset completeness largely determines the recall level of classification.

At 40 percent and 20 percent of the dataset scale, Recall drops to about 0.86 and 0.83, showing a clear performance degradation. This reveals that under data scarcity, the model is prone to losing critical feature information. It also shows that complex spatiotemporal modeling structures have a stronger dependency on training sample size. Although the model maintains a certain degree of robustness, it still struggles to fully demonstrate its advantages in small-sample scenarios.

Overall, the experimental results emphasize the importance of large-scale datasets for spatiotemporal fusion image classification models. Complete data not only improves the model's ability to learn dynamic features but also enhances adaptability to complex scenes. Therefore, reducing the reliance on data scale while ensuring performance becomes an important direction for further optimization. Approaches such as data augmentation or transfer learning can help compensate for performance loss caused by limited samples.

5. Conclusion

This study conducts a systematic investigation of deep image classification networks based on spatiotemporal feature fusion. From method design to performance validation, the proposed approach demonstrates significant advantages. By integrating spatial feature extraction and temporal dynamic modeling into the network, the model effectively captures semantic information across different dimensions. At the feature fusion stage, it achieves a unified representation of global and local information. This design not only overcomes the limitation of traditional methods that rely on a single dimension but also provides stronger discriminative ability for classification tasks in complex image scenarios. Experimental results show that the proposed model outperforms existing representative methods on multiple metrics, highlighting the importance and necessity of spatiotemporal feature fusion.

In terms of overall performance, the method improves classification accuracy while also showing outstanding stability and robustness. Under various conditions, the model demonstrates strong adaptability, maintaining good performance across different dataset scales and complex

backgrounds. This reflects the rationality of the structural design and its efficient support for learning large-scale data patterns. Through systematic comparisons and sensitivity experiments, the study further reveals the influence of dataset size, hyperparameter settings, and environmental conditions on model performance. These findings provide a useful reference for subsequent model optimization and application expansion.

On the application level, the proposed spatiotemporal fusion method has broad practical significance. In medical imaging, it can help track the progression of lesions over time and improve the scientific validity and reliability of clinical diagnosis. In intelligent transportation, the model better identifies risk states in dynamic environments, supporting traffic safety management and autonomous driving. In remote sensing, the method can be applied to classification tasks involving land cover changes and environmental monitoring, providing precise data support for urban planning and ecological protection. These application values fully demonstrate the contribution of the study to solving real-world problems.

In summary, this research not only proposes a new solution for spatiotemporal feature fusion from a theoretical perspective but also shows strong practical potential in applications. Future work may further explore the scalability of the method in larger and more complex scenarios, while considering integration with other deep learning techniques to enhance interpretability and generalization. By continuously optimizing the approach and expanding its application scope, this study is expected to provide a stronger technical foundation and innovative momentum for the development of artificial intelligence in medicine, transportation, safety, and remote sensing.

References

[1] Li J. and Zhao R., "Image Classification Based on A Spatiotemporal Convolutional Neural Network," Procedia Computer Science, vol. 243, pp. 356-363, 2024.

[2] Zhang Z., Zhang W., Meng Y., et al., "Satellite Image Time-Series Classification with Inception-Enhanced Temporal Attention Encoder," Remote Sensing, vol. 16, no. 23, p. 4579, 2024.

[3] Xu M., Zhou W., Shen X., et al., "Temporal-spatial cross attention network for recognizing imagined characters," Scientific Reports, vol. 14, no. 1, p. 15432, 2024.

[4] Shao P., Feng J., Zhang P., et al., "Interpretable spatial-temporal attention convolutional network for rainfall forecasting," Computers & Geosciences, vol. 185, p. 105535, 2024.

[5] Lee D., Li Y., Kim Y., et al., "Spiking transformer with spatial-temporal attention," Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 13948-13958, 2025.

[6] Ji S., Xu W., Yang M., and Yu K., "3D convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, 2012.

[7] Tran D., Bourdev L., Fergus R., Torresani L., and Paluri M., "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision, pp. 4489-4497, 2015.

[8] Simonyan K. and Zisserman A., "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems, vol. 27, 2014.

[9] Donahue J., Hendricks L. A., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K., and Darrell T., "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625-2634, 2015.

[10] Wang X., Girshick R., Gupta A., and He K., "Non-local neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803, 2018.

[11] Lian L., Li Y., Han S., Meng R., Wang S., and Wang M., "Artificial intelligence-based multiscale temporal modeling for anomaly detection in cloud services," arXiv preprint arXiv:2508.14503, 2025.

[12] Zhang X. and Wang Q., "EEG anomaly detection using temporal graph attention for clinical applications," Journal of Computer Technology and Software, vol. 4, no. 7, 2025.

[13] Liu Z., Ning J., Cao Y., Wei Y., Zhang Z., Lin S., and Hu H., "Video swin transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202-3211, 2022.

[14] Bertasius G., Wang H., and Torresani L., "Is space-time attention all you need for video understanding?," in International Conference on Machine Learning (ICML), vol. 2, no. 3, p. 4, July 2021.

[15] Feichtenhofer C., Fan H., Malik J., and He K., "Slowfast networks for video recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202-6211, 2019.

[16] Carreira J. and Zisserman A., "Quo vadis, action recognition? A new model and the kinetics dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308, 2017.

[17] Zou Y., Qi N., Deng Y., Xue Z., Gong M., and Zhang W., "Autonomous resource management in microservice systems via reinforcement learning," in 2025 8th International Conference on Computer Information Science and Application Technology (CISAT), pp. 991-995, IEEE, July 2025.

[18] Yao G., Liu H., and Dai L., "Multi-agent reinforcement learning for adaptive resource orchestration in cloud-native clusters," arXiv preprint arXiv:2508.10253, 2025.

[19] Lyu N., Wang Y., Cheng Z., Zhang Q., and Chen F., "Multi-Objective Adaptive Rate Limiting in Microservices Using Deep Reinforcement Learning," arXiv preprint arXiv:2511.03279, 2025.

[20] Sun Y., Meng R., Zhang R., Wu Q., and Wang H., "A Deep Q-Network Approach to Intelligent Cache Management in Dynamic Backend Environments," (preprint) 2025.

[21] Qi N., "Deep learning and NLP methods for unified summarization and structuring of electronic medical records," Transactions on Computational and Scientific Methods, vol. 4, no. 3, 2024.

[22] Pan D. W. A. S., "Dynamic Topic Evolution with Temporal Decay and Attention in Large Language Models," arXiv preprint arXiv:2510.10613, 2025.

[23] Simonyan K. and Zisserman A., "Two-stream convolutional networks for action recognition in videos," Advances in Neural Information Processing Systems, vol. 27, 2014.

[24] Peng S., Zhang X., Zhou L., and Wang P., "YOLO-CBD: Classroom Behavior Detection Method Based on Behavior Feature Extraction and Aggregation," Sensors, vol. 25, no. 10, p. 3073, 2025.

[25] Cui W., "Vision-Oriented Multi-Object Tracking via Transformer-Based Temporal and Attention Modeling," Transactions on Computational and Scientific Methods, vol. 4, no. 11, 2024.