Image Anal Stereol 2025;44:11-23 doi: 105566/ias.3361
Original Research Paper

IMAGE EXPERIENCE PREDICTION FOR HISTORIC DISTRICTS USING A CNN-TRANSFORMER FUSION MODEL

WEIJIA WANG[1,2], YOUPING TENG✉[,2], LU YAN[3], LONGWEI WU[2], YINYING YANG[4] AND ZIJIAN LUO[2]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou 310063, China; [2]School of Art and Archaeology, Hangzhou City University, Hangzhou 310015, China; [3]Graduate School of Science and Engineering, Chiba University, Chiba 263-8522, Japan; [4]School of Art and Design, Zhejiang Sci-Tech University, Hangzhou 310018, China

e-mail: w.weijia@zju.edu.cn; tengyp@hzcu.edu.cn; 18wd8305@student.gs.chiba-u.jp; wulw@hzcu.edu.cn; uninini_15@126.com; 32111017@stu.hzcu.edu.cn

(Received August 16, 2024; revised November 17, 2024; accepted November 22, 2024)

ABSTRACT

This study addresses fundamental challenges in historic district planning and design, specifically incorporating the emotional value of streetscape images into the design process. A deep learning-based sentiment analysis system was developed, utilising a convolution neural network (CNN) and transformer models to assess emotional tendencies and temporal states within images. The system employs a multi-view feature extraction framework, integrating VGG, ResNet CNNs, and the Swin Transformer model, resulting in a novel feature matrix. The attention mechanism and transfer learning strategy significantly enhance model accuracy in label recognition and classification. The primary contribution of this study is developing a novel multimodal fusion model, which markedly improves sentiment recognition accuracy and practical applicability. The application of this system to the Jiangnan Historic District underscores the enhanced appeal through the integration of emotional value. By identifying emotional tendencies in streetscape images, designers can make better-informed decisions that foster positive experiences. This research innovatively applies sentiment analysis to historic district design, highlighting the potential of the system to guide culturally sensitive and engaging urban planning. Our analysis of images from 12 Jiangnan historic districts demonstrated the efficiency of the system in aligning images with existing imaging libraries, offering valuable references and feedback. The results underscore the practical potential of deep learning in visual sentiment analysis and emphasize the significance of emotional value in enhancing experiences in historic districts. This study provides new insights and methodological support for planning and designing such areas.

Keywords: Convolutional Neural Network (CNN); Evaluation System; Historic Districts; Sentiment Analysis; Transformer Model.

INTRODUCTION

Over time, historic districts have increasingly assumed multifaceted roles in urban development. These districts, as custodians of urban culture, preserve rich historical narratives and significantly influence the emotional experiences of residents and visitors.

Consequently, the strategies employed in their design and preservation are pivotal for safeguarding the historical heritage and cultural dynamism of a city.

PROBLEM ANALYSIS OF THE CURRENT SITUATION OF HISTORIC DISTRICTS

The rejuvenation of historic districts in Jiangnan has spanned over four decades, beginning with the regeneration of Tunxi Old Street in 1983. Although this process has led to some physical environmental improvement, the prevailing urban regeneration model has introduced numerous issues.

DAMAGE TO SPATIAL TEXTURE

Driven by traditional urban renewal thinking, Jiangnan historic districts have adopted measures such as restoring old buildings, widening streets, opening blocked alleys, and improving public facilities. Consequently, these transformations extend beyond restoration, incorporating diverse new building materials and distinctive decorative elements, evolving into 'historic streets' that blend nostalgic charm with modern aesthetics. This synthesis caters to the refined tastes and consumption preferences of the upper class, reflecting a fusion of heritage preservation and contemporary luxury (Sanchez, 2023). While these new buildings satisfy

aesthetic tastes, they have marred the traditional look of these historic districts. Historic districts are envisioned as multifunctional spaces imbued with vibrancy, harmonising residents' living, production, interaction, and leisure activities. Nevertheless, as urbanisation accelerates, urban land resources have grown increasingly scarce (Giglio *et al.*, 2019), leading to a rapid widening of gaps in land rents, drastically affecting their original community functions and patterns of interpersonal interactions (Zhao *et al.*, 2018).

PURPOSE OF THE STUDY

This study aims to develop a sophisticated image sentiment analysis system using deep learning techniques to balance historic district preservation with modern urban development. By leveraging a hybrid model of CNNs and transformers, it seeks to enhance urban planners' and designers' understanding of public emotional responses, fostering empathy and cultural sensitivity in their design choices. This approach provides robust data support for preservation and renewal initiatives, offering real-time feedback and facilitating informed, historically sensitive urban regeneration strategies.

To balance historic preservation with urban development, this study uses both quantitative and qualitative measures to evaluate key structural variables: architectural integrity, streetscape cohesion, and pedestrian density. Quantitative scoring assesses the alignment of each variable with traditional or modern design requirements. Complementary qualitative insights from surveys capture local residents' and visitors' emotional responses to specific district changes. These approaches collectively provide a structured framework to determine the efficacy of design interventions in balancing cultural heritage with contemporary utility.

Building on this framework, the study utilises a deep learning model to categorise streetscape visual elements as positive or negative, measuring both sentiment intensity and diversity. By associating these emotional responses with specific design features—such as architectural details or vegetation—the model identifies elements that elicit favourable reactions. These insights are then translated into actionable design recommendations, enabling planners to refine historic district designs to better align with public sentiment, thereby enhancing engagement and positive emotional responses.

LITERATURE REVIEW

Sentiment analysis techniques traditionally applied to text have expanded to image and video analyses to explore cross-modal sentiment recognition. Research using similarity techniques for sentiment detection has shown promising results in recognising and analysing emotions in digital texts (Mozafari and Tahayori, 2019, Pratibha *et al.*, 2022). The application of CNN-Bi-LSTM models incorporating attention mechanisms in electroencephalogram (EEG)-based emotion recognition has validated these methods' maturity and efficacy (Huang *et al.*, 2023). In image sentiment analysis, numerous studies have utilised target detection algorithms to identify salient regions within images. Sentiment analysis is then conducted on both the salient target and original image separately using VGGNet, enhancing prediction accuracy through result fusion (Roshan *et al.*, 2024, Wu *et al.*, 2020). An RGBT (visible light and infrared) object tracking method based on multimodal hierarchical relationship modelling integrates multiple Transformer encoders and includes a dynamic component feature fusion module, significantly enhancing multimodal image feature aggregation and recognition accuracy, enabling dynamic importance assessment and contextual adaptation of each modality (Yao et al., 2024). A deep learning ensemble model for eye disease detection combines multi-layer CNN architectures such as VGG16, DenseNet201, and ResNet50, showing superior performance in medical image classification tasks (Jeny et al., 2023).

Experimental findings indicate that this fusion-based training approach outperforms traditional methods. Some studies have integrated visual self-attention mechanisms into CNN emotion classification frameworks, utilising salient targets in images as prior knowledge to optimise the visual attention learning area and address the limitations of self-attention mechanism in capturing emotional features (Song *et al.*, 2018, Thilagavathy *et al.*, 2023). Further innovations include an image sentiment analysis strategy that merges overall and local region embeddings for object localisation, employing deep neural networks to capture the sentiment attributes of localised regions, followed by classifiers for sentiment prediction (Cai *et al.*, 2019, Gupta *et al.*, 2023). Advances in CNNs enhanced with feature pyramids have facilitated multi-scale salient target emotion feature extraction by integrating saliency targets, facial recognition, and overall image analysis (Miao *et al.*, 2021). Sentiment information in images, often represented as high-level visual feature abstractions, is more readily interpreted through the subjects or attributes within the images (Xu *et al.*, 2014). Consequently, CNNs have transitioned from image subject recognition to sentiment analysis, enhancing their adaptability and accuracy in contexts rich with emotional and semantic content (Acheampong *et al.*, 2020, Zhang *et al.*, 2017).

Sentiment analysis of multimodal data utilises the integration of information from various modalities through associative and complementary methods to enhance sentiment classification (Chalasani *et al.*, 2020, Mahima *et al.*, 2021). Current fusion techniques for multimodal data are categorised into three types (Baltrušaitis *et al.*, 2018): feature-level (Poria *et al.*, 2016, Tao *et al.*, 2020), decision-level (Cao *et al.*, 2016, Yuqing *et al.*, 2019), and hybrid (Huang *et al.*, 2019, Truong and Lauw, 2019, Zhao *et al.*, 2019). Feature-level fusion combines features from different modalities early in the processing stage, decision-level fusion aggregates outcomes at the output stage, and hybrid fusion merges both to maximise information utilisation, with deep multimodal attention fusion frameworks exploring associations between images and text. For instance, the Deep Multimodal Attention Fusion (DMAF) framework (Huang *et al.*, 2019, Kumar *et al.*, 2021, You *et al.*, 2016b) has surpassed the Cross-modal Consistency Regression (CCR) model across several datasets. The CCR model enhanced with visual attention (CCR-V) (You *et al.*, 2017), the tree-structured Long Short-term Memory (LSTM) with visual attention (T-LSTM-E) (You *et al.*, 2016a), and the Tensor Fusion Network (TFN) model (Zadeh *et al.*, 2017) exemplify cutting-edge advancements in graphic fusion research. In computer vision, although CNNs remain dominant, there is increasing interest in integrating CNNs with self-attention mechanisms to develop novel model architectures. For instance, a standard transformer applied to image processing has attained moderate accuracy on medium-sized datasets compared to the traditional ResNet architecture. However, on larger datasets, the Vision Transformer (ViT) model (Dosovitskiy *et al.*, 2020) has achieved superior performance, equalling or surpassing state-of-the-art result in several image recognition benchmarks. These advancements underscore the significant potential of multimodal data fusion and deep learning techniques to drive further progress in sentiment analysis.

Song *et al.*, (2022) introduced a Multi-scale Fusion Network (CTMFNet), which integrates CNN and transformer mechanisms to enhance the semantic segmentation of remotely sensed urban scenes. This network leverages local and global contextual information through a multilayer densely connected network decoder, demonstrating superior accuracy over existing methods in practical scenarios. Notably, deep learning techniques, particularly CNNs and their integration with transformer models, have driven significant advancements in image classification, segmentation, and recognition (Espinoza *et al.*, 2023, Jayaswal *et al.*, 2024). A comprehensive review has highlighted these developments and the persisting challenges in the field, indicating a critical need for innovative deep models and computational systems to interpret image content more efficiently (Jiao and Zhao, 2019). In design evaluation, image classification is employed to assess compliance and aesthetic quality (Nasution *et al.*, 2023). Research into the aesthetic attributes of images across mixed multi-attribute datasets has advanced methodologies for assessing aesthetic quality (Bhushanam, 2023, Jin *et al.*, 2023, Jin *et al.*, 2019). Recent studies examining visual assessments of historic districts have evolved from traditional analyses of architectural styles and spatial layouts to modern approaches incorporating public participation and social dynamics (Middel *et al.*, 2019). By analysing visitor photography and perceptions, this research underscores the importance of preserving the authenticity and visual integrity of these areas (Naoi *et al.*, 2011).

This literature review highlights notable achievements in the utilisation of deep learning for image processing, advancements in sentiment analysis techniques, and real-world applications of image classification in design assessment.

These studies have enriched our understanding of visual assessment methodologies and exemplified the capabilities of deep learning models. Additionally, they demonstrate the extensive potential of sentiment analysis techniques in enhancing image analysis.

The current literature on sentiment analysis exhibits several limitations: it predominantly focuses on text-based methods, with limited incorporation of image-based analysis, thereby restricting comprehensive emotional representation across multimodal scenes. Studies investigating image learning in various historic district settings are also scarce, reducing the diversity of visual data utilised for model training and impacting adaptability. Traditional CNNs often struggle to manage complex visual contexts, and many models focus on localised features rather than adopting a holistic, integrated approach, thereby diminishing sentiment recognition accuracy. While sentiment analysis shows potential in design, practical applications in historic district contexts remain rare, leading to limited real-world validation. Furthermore, models trained on small, homogeneous datasets frequently fail to generalise across diverse historic settings, and temporal or spatial factors—such as time of day or seasonal variations—are often overlooked, restricting broader applicability. Future research should prioritise integrating multimodal data, diversifying visual data sources, enhancing model adaptability, and validating applications across diverse design contexts to better support sentiment analysis in historic districts.

MATERIALS AND METHODS

In this study, data were collected through field photography and Web access, resulting in a comprehensive dataset of 3,295 images related to historic districts (Shi *et al.*, 2023). Additionally, subjective evaluations of various street photographs from 12 historic districts (Table 1), encapsulating both positive and negative emotions, were conducted through interviews with 426 tourists. Each image in the dataset was meticulously labelled to facilitate detailed analysis.

Table 1: *Summary of the results of information on 12 popular historic districts in Jiangnan. Street-level images of these districts were acquired for the field survey.*

Serial num.	**Block (between streets)**	**Spatial texture**	**District atmosphere**
1	Huangshan Tunxi Old Street	The fishbone pattern is complete, featuring numerous Huizhou-style buildings with high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The branch alleys feature many original residents, old shops, teahouses, and daily amenities such as vegetable markets and hardware shops, which contribute to a strong sense of local life. Conversely, the main street is dominated by tourists, hotels, commercial establishments, creating a distinctly commercial environment.
2	Suzhou Pingjiang Road Districts	The checkerboard pattern is complete, featuring numerous Jiangnan buildings with high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The branch alleys are characterised by numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, creating a strong sense of local life. In contrast, the main street is frequented by many tourists and features an abundance of cafes, hotels, and shops, fostering a distinctly commercial environment.
3	Suzhou Shantang Old Street	The one-river-and-two-streets pattern is complete, featuring numerous Jiangnan buildings with high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The branch alleys host numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, creating a strong sense of local life. Conversely, the main street is frequented by tourists and features an abundance of cafes, hotels, and shops, creating a distinctly commercial environment.
4	Shaoxing Houttuyniau Historic Quarter	The grid pattern is complete, featuring numerous Huizhou-style buildings with high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The branch alleys feature numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, contributing to a strong sense of local life. In contrast, there are few tourists, cafes, hotels, and shops, resulting in a weak commercial atmosphere.
5	Lanxi Tianfu Mountain Districts	The grid pattern is complete, featuring numerous Huizhou-style buildings with high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The branch alleys are characterised by numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, creating a strong sense of local life. Conversely, there are few tourists, cafes, hotels, and shops, resulting in a weak commercial atmosphere.
6	Hangzhou Nan Song Royal Street	The fishbone pattern is complete, integrating Jiangnan and Chinese–Western architecture, characterised by high spatial integration, unobstructed streets and lanes with good accessibility.	The branch alleys host numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, fostering a strong sense of local life. The area attracts numerous tourists and is populated with cafes, express hotels, speciality lodgings, and shops, reflecting a strong commercial presence.
7	Shanghai Tianzifang	The field pattern is complete, featuring numerous Shikumen buildings, high interface similarity, high spatial integration, and unobstructed streets and lanes with good accessibility.	The district exhibit sparse original residents, few old shops, a lack of daily amenities such as vegetable markets, hardware shops, and night markets. This contributes to a weak sense of local life. In contrast, there are more tourists, cafes, hotels, and shops, creating a strong commercial atmosphere.
8	Shanghai Xintiandi (shopping, eating, and entertainment district of Shanghai)	The grid pattern is complete, featuring numerous Shikumen buildings, high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The area has few original residents, fewer old shops, and lacks daily amenities such as vegetable markets, hardware shops and night markets, leading to a weak sense of local life. Conversely, it attracts numerous tourists and features many cafes, hotels, and shops, resulting in a strong commercial atmosphere.
9	Hangzhou Qiaoxi Districts	The river–street–alley pattern is complete, incorporating numerous Jiangnan buildings, high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The branch alleys host numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, fostering a strong sense of local life. Additionally, there are many tourists, hotels, and shops, contributing to a strong commercial atmosphere.

10	Hangzhou Wuhe Historic District	The street–river–street pattern is complete, featuring numerous Huizhou-style buildings, high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The area hosts numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, fostering a strong sense of local life. Conversely, the presence of few tourists, few cafes, hotels, and shops, contributes to a weak commercial atmosphere.
11	Shaoxing Cangqiao Zhijie	Fishbone layout is well-defined, featuring numerous Huizhou-style buildings, high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	The area hosts numerous original residents, old shops, teahouses, and daily amenities such as vegetable markets, hardware shops, and night markets, fostering a strong sense of local life. In contrast, there are few tourists, cafes, hotels, and shops, resulting in a weak commercial atmosphere.
12	Huangshan Liyang Old Street	The area features numerous Huizhou-style buildings with high interface similarity, high spatial integration, and smooth, accessible streets and lanes.	There are few original residents, a lack of daily habitats, such as old shops, vegetable markets, hardware shops, and night markets, and a weak sense of life. There are many tourists, cafes, hotels, and shops, and a strong commercial atmosphere.

STUDY AREA

The Jiangnan region generally encompasses areas south of the Yangtze River, primarily southern Jiangsu, northern Zhejiang, and southern Anhui in eastern China. Known for its warm and humid climate, extensive network of waterways, and fertile land, Jiangnan is renowned as the "Land of Fish and Rice." The historical districts in Jiangnan cities are intricately integrated with the natural environment, featuring interconnected waterways and lush greenery that exemplify the distinctive charm of water towns in the region. These districts preserve rich historical and cultural information, documenting the development and transformation of Jiangnan, and actively protect the ecological environment and maintain urban ecological balance. Characterised by exquisite architectural art and traditional garden landscapes, Jiangnan's historical urban districts exhibit high aesthetic value and have become significant tourist attractions, drawing numerous domestic and international visitors.

In the geographic map, the primary historical districts are marked with red dots using GIS tools (Fig. 1), predominantly concentrated in the core area of the Yangtze River Delta. This distribution underscores the unique integration of human and natural landscapes in the Jiangnan region.

EXPERIMENTAL SETUP

Experiments utilised high-performance computing resources, primarily Google Colab, for efficient model training and evaluation. Essential Python packages were imported and the dataset was divided into training, validation, and testing subsets. PyTorch DataLoader managed batch processing and data shuffling. The setup prioritised GPUs using CUDA technology, defaulting to CPUs if GPUs were unavailable, ensuring optimal resource utilisation. All models were trained and tested in a standardised environment with carefully selected hyperparameters. The dataset was segmented into training and testing portions for performance evaluation (Garcea *et al.*, 2023, Xia *et al.*, 2017).

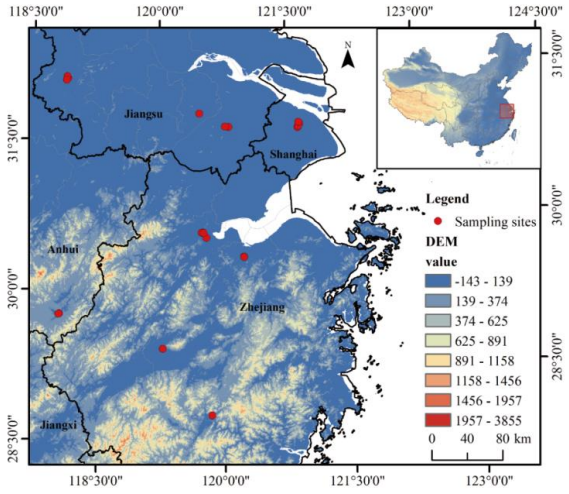

Fig. 1. *Distribution of major historical districts in the Jiangnan Region.*

Street photographs from 12 typical historic districts were selected and pre-processed through resizing and normalisation to ensure high-quality data input into the model. The dataset was split into training (80%) and testing (20%) sets to optimise model training and test evaluation precision.

SENTIMENT CLASSIFICATION DATASET

Participants in the evaluation classified each photograph as either 'positive' or 'negative' based on their emotional reactions. To meet the input specifications of the model, all collected street photographs underwent initial pre-processing, including cropping and standardisation of dimensions. Each photograph was then assigned a sentiment label derived from a consensus of tourist evaluations. For label accuracy and consistency, a

photograph was included in the final dataset only if it achieved > 75% agreement in tourist ratings.

SYSTEM FRAMEWORK MODELLING INNOVATION INTEGRATION

The advanced image emotion learning and prediction system developed in this study employs a hybrid model architecture incorporating classical CNNs and a sophisticated transformer model (Khan *et al.*, 2019, Zunair and Hamza, 2020). This integrated deep learning model, named 'Big Model', aims to improve the accuracy of emotion classification for images of historic districts by extracting and amalgamating multiple visual features. As illustrated in Fig. 2, the model synergistically combines three state-of-the-art deep-learning architectures: VGG13, ResNet18, and Swin Transformer. Each architecture offers distinct advantages: VGG13 excels in capturing texture details, ResNet18 in discerning global structural information, and the Swin Transformer in identifying dynamic relationships (Tu *et al.*, 2022). Each base model independently extracts features from the input images, capturing diverse information from various viewpoints and scales. These extracted features are then merged through a transformer encoder layer (Shang *et al.*, 2023), optimising their complementary strengths. Finally, the integrated feature vectors are processed using a fully connected layer to classify and output the emotion category of the image. The training process for the fusion model was structured into three sequential steps: pre-processing, model fine-tuning, and end-to-end training (Zhang *et al.*, 2024).

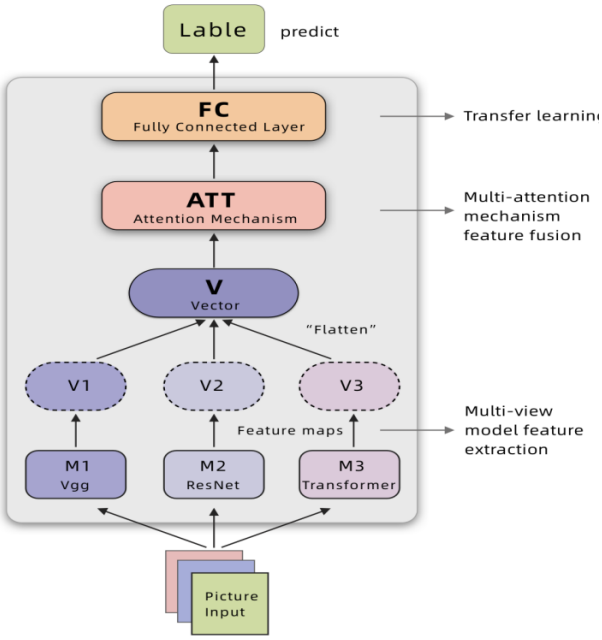

Fig. 2. *The framework of Big Model structure*.

IMAGE PREDICTION MODEL

The image prediction model (Fig. 3) utilises cosine similarity to assess the resemblance between a specific query vector and a set of target vectors, aiming to identify the top k vectors that most closely match the query vector (Sejal *et al.*, 2016). The process begins by calculating similarity scores between the query vector and each vector in the target set using a cosine similarity function:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

These scores are sorted to identify the k vectors exhibiting the highest similarity. The model selects the k indices with the highest similarity scores and their corresponding values, efficiently identifying the vector most resembling the query vector. This method benefits applications such as recommender systems, image recognition, and other scenarios requiring rapid identification of similar items (Islam *et al.*, 2024, Liu *et al.*, 2022). Upon training completion, the trained 'Model Big' was employed to predict the sentiment of images from historic districts. Following pre-processing, the image was input into the model, which classified the emotional category and output the corresponding prediction probability. The model determines whether an image is likely to portray a positive or negative emotion based on prediction probability.

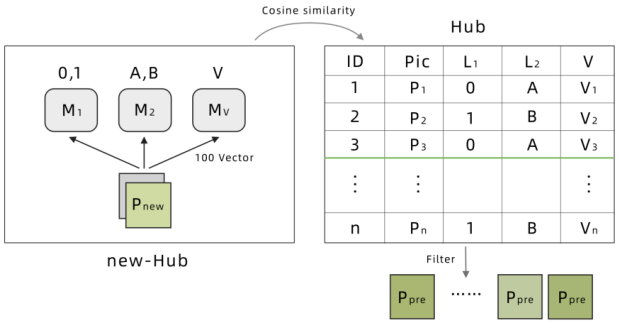

Fig. 3. *The framework of image prediction model*.

RESULTS

PERFORMANCE COMPARISON OF DEEP LEARNING MODELS

We conducted tests on 1,480 graphs, producing classification report detailed in Tables 2 to 5. We evaluated the performance of three deep learning models—CNN, ResNet, and Swin Transformer—and their combined fusion model on a classification task with 1,480

images. The Swin Transformer and fusion models exhibited high accuracy and robust performance, outperforming the CNN and ResNet models. The CNN model achieved 88.9% accuracy, with precision and recall varying by category. The ResNet model attained 84.4% accuracy, with high precision but varied recall. The Swin Transformer excelled with 97.8% accuracy, demonstrating high precision and recall across both categories. The fusion model achieved 97.8% accuracy, equal to the Swin Transformer, highlighting the benefits of a multi-model fusion strategy for complex classification tasks requiring high accuracy and reliability. This combination effectively captured the complex visual features of the Jiangnan Historic District for accurate sentiment analysis.

Table 2. *CNN model classification report*:

	precision	**recall**	**f1-score**	**support**
0	0.842	0.889	0.865	592
1	0.923	0.889	0.906	888
accuracy			0.889	1480
macro avg	0.883	0.889	0.885	1480
weighted avg	0.891	0.889	0.889	1480

Table 3. *ResNet classification report*:

	precision	**recall**	**f1-score**	**support**
0	0.720	1.000	0.837	592
1	0.923	0.889	0.851	888
accuracy			0.844	1480
macro avg	0.860	0.870	0.844	1480
weighted avg	0.888	0.844	0.846	1480

Table 4. *Swin Transformer classification report*.

	precision	**recall**	**f1-score**	**support**
0	0.947	1.000	0.973	592
1	1.000	0.963	0.981	888
accuracy			0.978	1480
macro avg	0.974	0.981	0.977	1480
weighted avg	0.979	0.978	0.978	1480

Table 5. *Three-model fusion classification report*:

	precision	**recall**	**f1-score**	**support**
0	1.000	0.955	0.977	724
1	0.958	1.000	0.979	756
accuracy			0.978	1480
macro avg	0.979	0.977	0.978	1480
weighted avg	0.979	0.978	0.978	1480

CONSTRUCTION OF A POSITIVE-NEGATIVE MODELLING SYSTEM

We evaluated the positive-negative model over five training cycles, processing 1,620 images per cycle. The performance of the model steadily improved, evidenced by a reduction in loss and an increase in test accuracy and stability. Detailed results for each epoch are presented in Fig. 4. In Epoch 5, the initial loss was 0.560, and the end loss was reduced to 0.436. The test accuracy peaked at 88.8%, and the average loss decreased to 0.378. The second test maintained an accuracy of 71.4% with an average loss of 0.592.

These results demonstrate a continuous reduction in loss and consistent improvement in accuracy across the training epochs, reflecting gradual model optimisation and enhanced learning efficiency. These findings highlight the effectiveness of the model in image classification tasks, showcasing its ability to adapt and improve over time.

To thoroughly assess the performance of our binary classification model, we employed the confusion matrix function from the sklearn. metrics library, which computes a matrix based on actual labels (label_array) and predicted labels (predict_array). The confusion matrix proved invaluable for evaluating the performance of our classification model across different categories (Fig. 5).

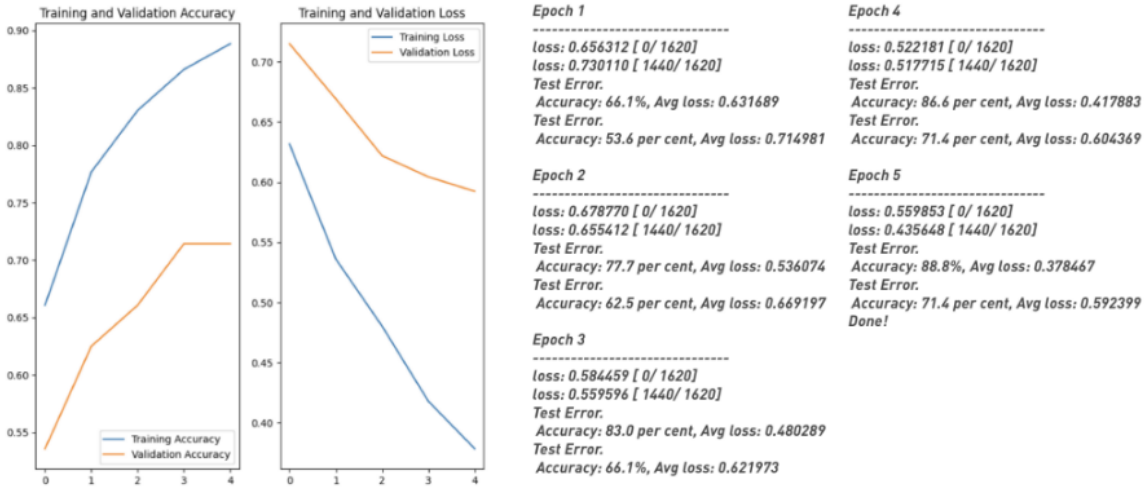

Fig. 4. *Five training epoch of the positive–negative modelling system.*

Positive–Negative Modelling

	precision	recall	f1-score	support
0	0.667	0.769	0.714	751
1	0.769	0.667	0.714	869
accuracy			0.714	1620
macro avg	0.718	0.718	0.714	1620
weighted avg	0.722	0.714	0.714	1620

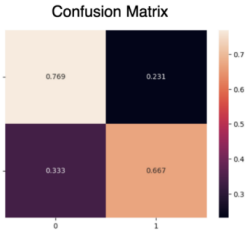

Fig. 5. *Positive–negative modelling classification report and confusion matrix*.

The confusion matrix visually and numerically represents model performance, with rows indicating actual categories and columns indicating predicted categories. This matrix effectively demonstrates the capability of the model to distinguish between categories, highlighting areas of strong performance and those requiring improvements.

APPLICATION AND PERFORMANCE OF THE PREDICTIVE MODEL

In this study, we employed the model to predict a single nighttime image from the Qiaoxi Historic District (Fig. 6). Located west of Gongchen Bridge, along the Hangzhou section of the Grand Canal, this district epitomises canal transport culture. Presently, Qiaoxi district preserves a blend of heritage sites and modern establishments, such as Huichun Hall, courtyards, piers, pavilions, museums, shopping centres, and cafes, creating a rich commercial atmosphere.

In our analysis, we selected images from the galleries for testing (Table 6 and Fig. 7). The results demonstrate that the model is highly effective in identifying images similar to the input. The most similar image was the same nighttime photo of image No.20, used as the input. The second most similar image was a nighttime scene from image No.9, and the third was another view from image No.15. Despite foreground similarities, the system effectively recognised distinct differences in the scenes, including modern residential buildings in the background of the third image.

Fig. 6. *Photo of nighttime image from the Qiaoxi Historic District*.

The model recommended six images based on their high visual similarity and shared positive emotional tones, demonstrating its ability to efficiently retrieve and recognise images that align with the input in style and mood. It accurately captures emotional tendencies, particularly positive emotions, which is valuable for emotion analysis in historic districts or temporal settings. This capability aids urban planning and historic preservation projects by enabling designers to predict realistic scenarios and assess emotional states before renovations, thus assisting decision-makers in effectively utilising visual and emotional information.

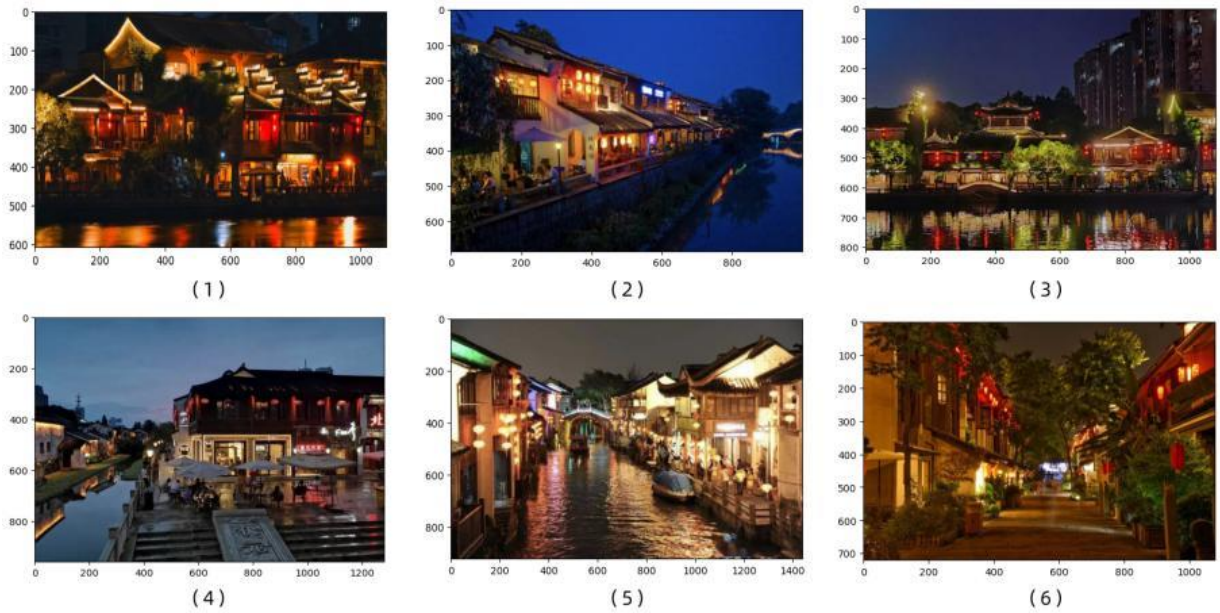

Fig. 7. *Filter out the corresponding images*.

Table 6. *List of filter out the corresponding images*.

	name	**l_emotion**	**l_time**	**file_path**
20	Qiaoxi	Positive	Night	/content/QX2.jpeg
9	Wuhe	Positive	Night	/content/WH2.jpeg
15	Qiaoxi District	Positive	Night	/content/QX11.jpeg
11	Pingjiang Road	Positive	Night	/content/PJ8.jpeg
16	Tunxi Old Street	Positive	Night	/content/TX14.jpeg
6	Southern Song Imperial Street	Positive	Night	/content/NSYJ6.jpeg
7	Cambridge and Kurama	Positive	Night	/content/CQ12.jpeg
5	Tunxi Old Street	Positive	Night	/content/TX5.jpeg
14	Mountain Pond	Positive	Night	/content/ST7.jpeg
4	Punjab, Province of Pakistan	Positive	Night	/content/WH23.jpeg

DISCUSSION

HOMOGENISATION ANALYSIS OF HISTORIC DISTRICTS

In restoring historic districts, renovation efforts often prioritise operational convenience or cost control. Typically, these projects preserve a few iconic elements, such as ancient trees, courtyards, and stone bridges, while introducing modern public art. However, there is a tendency to demolish many ordinary old buildings and redesign streets and lanes to meet spatial quality evaluation indices such as integration, accessibility, and comprehensibility. Unfortunately, these aggressive renovation practices can disrupt the original spatial texture of districts and sever the social bonds these spaces embody, thus erasing the collective memory of the area. Additionally, the transformation often involves replacing everyday staples, such as vegetable markets, teahouses, hardware shops, and night markets, with upscale venues featuring international brands and fashion symbols. This shift significantly alters the once vibrant, bustling atmosphere filled with local colour, distancing it from its historical roots.

Utilising deep learning models such as CNNs and transformers to analyse image data from historic districts allows us to identify visual elements commonly associated with specific emotional labels. This analysis elucidates the

intricate relationship between visual elements and emotional responses and aids in understanding the underlying causes of visual and emotional homogeneity across different districts.

ADVANTAGES OF FUSION MODELLING

Merging three deep learning architectures, specifically ResNet, VGG, and Swin Transformer, into a single integrated model offers multiple advantages for image recognition and processing tasks.

Comprehensive feature extraction capability: ResNet and VGG models effectively capture local image features, such as edges, textures, and shapes, which are crucial for detail comprehension. Conversely, the Swin Transformer excels in capturing global dependencies via a self-attention mechanism. The integration of both local and global features enhances the comprehensive understanding of image content.

Efficiency and accuracy trade-off: In practical applications, a balance must be maintained between the computational efficiency of a model and its prediction performance. By integrating the models, the high efficiency of ResNet, the robust feature extraction of VGG, and the global understanding capability of the Swin Transformer can be leveraged to achieve high-precision prediction at relatively low computational costs.

Enhanced performance in complex scenes: Complex scenes, such as images of historic districts, encompass diverse information, including various architectural styles and complex background elements. The three-model fusion approach improves handling of complex data and enhances the accuracy of classification, recognition, and other tasks by combining their distinct features.

RESEARCH LIMITATIONS AND FUTURE PROSPECTS

The proposed model effectively integrates tourists' subjective ratings with automated image analysis for emotion classification. However, it encounters challenges in capturing the complexity of emotional states. To address this, more advanced models and algorithm incorporating multidimensional data, such as textual comments and social media feedback, are being developed. Temporal state classification is heavily influenced by lighting and sky colour, necessitating robust feature recognition techniques and the inclusion of multimodal data, such as weather information, to improve accuracy. Future research should integrate multiple data sources, including text and sound, with image data to enhance the comprehensiveness of sentiment analysis. Ongoing innovations in model structures, including leveraging advanced techniques such as GANs and self-supervised learning, will enhance model capabilities and generalisability.

In conclusion, the experimental results provide significant insights into the performance, limitations, and potential applications of the multimodal sentiment analysis model developed for historic district imagery. First, the fusion model, integrating CNN, ResNet, and Swin Transformer architectures, offers considerable advantages over single-model approaches, particularly in handling the complex visual contexts typical of historic districts. This superiority is evidenced by its enhanced accuracy, precision, and recall, highlighting the efficacy of multimodal integration in capturing both local detail and global dependencies. Detailed sentiment classification analysis reveal that the model excels in distinguishing between positive and negative emotions, as confirmed by the confusion matrix, suggesting that the ability of the Swin Transformer to capture nuanced emotional cues is instrumental in improving classification accuracy. Further analysis highlights the impact of specific visual features—such as architectural styles and cultural symbols—on sentiment outcomes, offering practical insights into the ways these elements shape public emotional responses. This capability renders the model a valuable tool for sentiment-driven urban design, allowing designers to anticipate and incorporate elements that evoke positive reactions. Nonetheless, limitations persist, such as decreased model performance in extreme settings including low-light or crowded scenes. Addressing these limitations in future studies—by expanding dataset diversity and enhancing image preprocessing techniques—could improve adaptability. These findings underscore the potential of multimodal sentiment analysis to inform culturally sensitive design and suggest pathways for future optimisation and broader application.

CONCLUSION

This study presents an image-emotion learning system employing deep-learning techniques tailored to classify emotions and temporal states in street photographs of historic districts. By integrating CNN and transformer models with subjective tourist evaluations, the model achieved over 80% classification accuracy across multiple dimensions. These findings affirm the efficacy of hybrid models in image sentiment analysis and underscore the potential of merging human subjective perception with machine-vision analysis. This system provides urban planners and designers with a novel tool for assessing the emotional impacts of historic districts. The emotion-based assessment approach offers robust data for creating emotionally appealing and resonant urban spaces. Additionally, analysing the emotional tendencies in images of historic districts enhances understanding of public emotional reactions to different historic environments, which is essential for humane and

sensitive urban cultural heritage preservation and restoration.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Zhejiang University and Hangzhou City University for their invaluable support and resources throughout this research.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author.

FUNDING

This research was funded by Major Humanities and Social Science Research projects of universities in Zhejiang Province: Study on Red Cultural Design Empowerment of historic blocks in Yangtze River Delta in the New Era 2023GH0281 and Beijing-Hangzhou Grand Canal Culture Research Institute: Research on regenerative conservation of historic districts along the Beijing-Hangzhou Grand Canal: JHY-2022YB05.

DATA AVAILABILITY

The datasets generated and analysed during this study were collected and photographed by the authors. User evaluations were conducted with participants' consent. These data are available from the corresponding author upon reasonable request.

REFERENCES

Acheampong FA, Wenyu C, Nunoo-Mensah H (2020). Text-based emotion detection: Advances, challenges, and opportunities. Eng Rep 2:e12189. doi: 10.1002/eng2.12189.

Baltrušaitis T, Ahuja C, Morency L-P (2018). Multimodal machine learning: A survey and taxonomy. IEEE T Pattern Anal 41:423-43. doi: 10.1109/TPAMI.2018.2798607.

Bhushanam PN (2023). 2023 2nd International Conference on Edge Computing and Applications (ICECAA); IEEE, 563-7. doi: 10.1109/ICECAA58104.2023.10212227.

Cai G, He X, Chu Y (2019). Visual sentiment analysis by combining global and local regions of image. J Comput Appl 39:2181. doi: 10.11772/j.issn.1001-9081.2018122452.

Cao D, Ji R, Lin D, Li S (2016). A cross-media public sentiment analysis system for microblog. Multimedia Syst 22:479-86. doi: 10.1007/s00530-014-0407-8.

Chalasani N, Gurujala SS, Kota SSS, Nishitha SNT, Kiran JS (2020). 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS); IEEE, 869-75. doi: 10.1109/ICISS49785.2020.9316040.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. doi: 10.48550/arXiv.2010.11929.

Espinoza S, Aguilera C, Rojas L, Campos PG (2023). Analysis of fruit images with deep learning: A systematic literature review and future directions. IEEE Access. doi: 10.1109/ACCESS.2023.3345789.

Garcea F, Serra A, Lamberti F, Morra L (2023). Data augmentation for medical imaging: A systematic literature review. Comput Biol Med 152:106391. doi: 10.1016/j.compbiomed.2022.106391.

Giglio S, Bertacchini F, Bilotta E, Pantano P (2019). Using social media to identify tourism attractiveness in six italian cities. Tourism Manage 72:306-12. doi: 10.1016/j.tourman.2018.12.007.

Gupta SK, Alemran A, Singh P, Khang A, Dixit CK, Haralayya B (2023). 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC); IEEE, 1-6. doi: 10.1109/ICRTEC56977.2023.10111885.

Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019). Image–text sentiment analysis via deep multimodal attentive fusion. Knowl-Based Syst 167:26-37. doi: 10.1016/j.knosys.2019.01.019.

Huang Y, Yi Y, Chen Q, Li H, Feng S, Zhou S, Zhang Z, Liu C, Li J, Lu Q (2023). Analysis of eeg features and study of automatic classification in first-episode and drug-naïve patients with major depressive disorder. BMC Psychiatry 23:832. doi: 10.1186/s12888-023-05349-9.

Islam M, Zunair H, Mohammed N (2024). Cossif: Cosine similarity-based image filtering to overcome low inter-class variation in synthetic medical image datasets. Comput Biol Med 172:108317. doi: 10.1016/j.compbiomed.2024.108317.

Jayaswal V, Ji S, Singh V, Singh Y, Tiwari V (2024). 2024 2nd International Conference on Disruptive Technologies (ICDT); IEEE, 1428-33. doi: 10.1109/ICDT61202.2024.10489470.

Jeny AA, Junayed MS, Islam MB (2023). Deep neural network-based ensemble model for eye diseases detection and classification. Image Anal Stereol 42(2):77-91. doi: 10.5566/ias.2857.

Jiao L, Zhao J (2019). A survey on the new generation of deep learning in image processing. IEEE Access 7:172231-63. doi: 10.1109/ACCESS.2019.2956508.

Jin X, Wu L, Zhao G, Li X, Zhang X, Ge S, Zou D, Zhou B, Zhou X (2019). Proceedings of the 27th ACM international conference on multimedia; 311-9. doi: 10.1145/3343031.3350970.

Jin X, Li Y, Zhou W, Zhou X, Yang H (2023). 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW); IEEE, 359-64. doi: 10.1109/ICMEW59549.2023.00068.

Khan MA, Javed MY, Sharif M, Saba T, Rehman A (2019). 2019 international conference on computer and information sciences (ICCIS); IEEE, 1-7. doi: 10.1109/ICCISci.2019.8716400.

Kumar HSH, Gowramma YP, Manjula SH, Anil D, Smitha N (2021). 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV); IEEE, 1332-7. doi: 10.1109/ICICV50876.2021.9388522.

Liu H, Zheng C, Li D, Zhang Z, Lin K, Shen X, Xiong NN, Wang J (2022). Multi-perspective social recommendation method with graph representation learning. Neurocomputing 468:469-81. doi: 10.1016/j.neucom.2021.10.050.

Mahima MA, Patel NC, Ravichandran S, Aishwarya N, Maradithaya S (2021). 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES); IEEE, 1-6. doi: 10.1109/ICSES52305.2021.9633843.

Miao Y, Lei Q, Zhang W, Wen Y (2021). Research on image emotional analysis of multi visual object fusion. Comput Appl Res 38:1250-5. doi: 10.19734/j.issn.1001-3695.2020.02.0087.

Middel A, Lukasczyk J, Zakrzewski S, Arnold M, Maciejewski R (2019). Urban form and composition of street canyons: A human-centric big data and deep learning approach. Landscape Urban Plan 183:122-32. doi: 10.1016/j.landurbplan.2018.12.001.

Mozafari F, Tahayori H (2019). 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS); IEEE, 1-5. doi: 10.1109/CFIS.2019.8692152.

Naoi T, Yamada T, Iijima S, Kumazawa T (2011). Applying the caption evaluation method to studies of visitors' evaluation of historical districts. Tourism Manage 32:1061-74. doi: 10.1016/j.tourman.2010.09.005.

Nasution FBB, Nasution N, Hasan MA (2023). 2023 International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE); IEEE, 90-5. doi: 10.1109/ICCTEIE60099.2023.10366623.

Poria S, Chaturvedi I, Cambria E, Hussain A (2016). 2016 IEEE 16th international conference on data mining (ICDM); IEEE, 439-48. doi: 10.1109/ICDM.2016.0055.

Pratibha, Khurana M, Kaur G, Kaur A (2022). 2022 10th international conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO); IEEE, 1-5. doi: 10.1109/ICRITO56286.2022.9964967.

Roshan M, Rawat M, Aryan K, Lyakso E, Mekala AM, Ruban N (2024). Linguistic based emotion analysis using softmax over time attention mechanism. Plos One 19:e0301336. doi: 10.1371/journal.pone.0301336.

Sanchez TW (2023). Planning on the verge of ai, or ai on the verge of planning. Urban Sci 7:70. doi: 10.3390/urbansci7030070.

Sejal D, Ganeshsingh T, Venugopal KR, Iyengar SS, Patnaik LM (2016). Image recommendation based on anova cosine similarity. Procedia Comput Sci 89:562-7. doi: 10.1016/j.procs.2016.06.091.

Shang J, Gao M, Li Q, Pan J, Zou G, Jeon G (2023). Hybrid-scale hierarchical transformer for remote sensing image super-resolution. Remote Sens 15:3442. doi: 10.3390/rs15133442.

Shi L, Luo J, Zhu C, Kou F, Cheng G, Liu X (2023). A survey on cross-media search based on user intention understanding in social networks. Inform Fusion 91:566-81. doi: 10.1016/j.inffus.2022.11.017.

Song K, Yao T, Ling Q, Mei T (2018). Boosting image sentiment analysis with visual attention. Neurocomputing 312:218-28. doi: 10.1016/j.neucom.2018.05.104.

Song P, Li J, An Z, Fan H, Fan L (2022). Ctmfnet: Cnn and transformer multiscale fusion network of remote sensing urban scene imagery. IEEE T Geosci Remote 61:1-14. doi: 10.1109/TGRS.2022.3232143.

Tao F, Peng W, Qi C (2020). The research of sentiment recognition of online users based on dnns multimodal fusion. J Inform Resour Manag 10:39-48.

Thilagavathy A, Suresh KH, Chowdary KT, Tejash M, Chakradhar VL (2023). 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA); IEEE, 1-6. doi: 10.1109/ICIDCA56705.2023.10099672.

Truong Q-T, Lauw HW (2019). Proceedings of the AAAI conference on artificial intelligence; 305-12. doi: 10.1609/aaai.v33i01.3301305.

Tu J, Mei G, Ma Z, Piccialli F (2022). Swcgan: Generative adversarial network combining swin transformer and cnn for remote sensing image super-resolution. IEEE J Sel Top Appl Earth Obs Remote Sens 15:5662-73. doi: 10.1109/JSTARS.2022.3190322.

Wu L, Qi M, Jian M, Zhang H (2020). Visual sentiment analysis by combining global and local information. Neural Process Lett 51:2063-75. doi: 10.1007/s11063-019-10027-7.

Xia G-S, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L, Lu X (2017). Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE T Geosci Remote 55:3965-81. doi: 10.1109/TGRS.2017.2685945.

Xu C, Cetintas S, Lee K, Li L (2014). Visual sentiment prediction with deep convolutional neural networks. arXiv preprint arXiv:14115731.

Yao R, Qiu J, Zhou Y, Shao Z, Liu B, Zhao J, Zhu H (2024). Visible and infrared object tracking based on multimodal hierarchical relationship modeling. Image Anal Stereol 43(1):41-51.

You Q, Cao L, Jin H, Luo J (2016a). Proceedings of the 24th ACM international conference on Multimedia; 1008-17. doi: 10.1145/2964284.2964288.

You Q, Luo J, Jin H, Yang J (2016b). Proceedings of the Ninth ACM international conference on Web search and data mining; 13-22. doi: 10.1145/2835776.2835779.

You Q, Jin H, Luo J (2017). Proceedings of the AAAI conference on artificial intelligence; doi: 10.1609/aaai.v31i1.10501.

Yuqing M, Junhong W, Tonglai L (2019). Joint vissual-textual approach for microblog sentiment analysis. Comput Eng Design 40:1099-105.

Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:170707250. doi: 10.48550/arXiv.1707.07250.

Zhang W, Tan Z, Lv Q, Li J, Zhu B, Liu Y (2024). An efficient hybrid cnn-transformer approach for remote sensing super-resolution. Remote Sens 16:880. doi: 10.3390/rs16050880.

Zhang Y, Wang Y, Jin J, Wang X (2017). Sparse bayesian learning for obtaining sparsity of eeg frequency bands based feature vectors in motor imagery classification. Int J Neural Syst 27:1650032. doi: 10.1142/S0129065716500325.

Zhao W, Wang H, Li Y, Wang Z (2018). The history, current dilemmas and coordination mechanism of urban settlements in china: Based on social and spatial perspectives. Urban Plan J 20-8.

Zhao Z, Zhu H, Xue Z, Liu Z, Tian J, Chua MCH, Liu M (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. Inform Process Manag 56:102097. doi: 10.1016/j.ipm.2019.102097.

Zunair H, Hamza AB (2020). Melanoma detection using adversarial training and deep transfer learning. Phys Med Biol 65:135005. doi: 10.1088/1361-6560/ab86d3.