



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Hierarchical fusion of multi-scale features with transformer for crime scene trace identification

Jiacheng Zhu

DOI: [10.1504/IJICT.2026.10076763](https://doi.org/10.1504/IJICT.2026.10076763)

Article History:

Received:	07 November 2025
Last revised:	07 December 2025
Accepted:	09 December 2025
Published online:	18 March 2026

Hierarchical fusion of multi-scale features with transformer for crime scene trace identification

Jiacheng Zhu

Department of Forensic Science and Technology,
Henan Police College,
Zhengzhou, 450000, China
Email: zjc@hnp.edu.cn

Abstract: To address the technical challenge of low identification accuracy for crime scene traces in complex environments, this paper proposes a multi-branch feature fusion model based on transformers. Traditional methods rely on manually selected features, while existing deep learning models struggle to balance trace-level details with global structural patterns. To clarify this point, these global structural patterns indeed often encompass critical spatial distribution characteristics of trace features. The relative positions, densities, and geometric relationships of features are fundamental to achieving accurate and robust trace matching. This study constructs four specialised branches – fingerprints, tool marks, footprints, and biological traces – to extract distinctive features from each trace type. It further designs cross-branch interaction and adaptive fusion mechanisms to achieve effective integration of multi-source trace information. Experiments on public datasets demonstrate that our method achieves an average identification accuracy of 95.7%, surpassing mainstream models by over 6%. This significantly enhances trace recognition capabilities in complex scenarios involving blurred or fragmented evidence.

Keywords: crime scene traces; transformer model; multi-branch feature fusion; trace identification.

Reference to this paper should be made as follows: Zhu, J. (2026) 'Hierarchical fusion of multi-scale features with transformer for crime scene trace identification', *Int. J. Information and Communication Technology*, Vol. 27, No. 23, pp.17–34.

Biographical notes: Jiacheng Zhu is working in the Department of Forensic Science and Technology, Henan Police College. China. He received his Bachelor's degree from the Criminal Investigation Police University of China, 2016, and Master's degree Henan from the University in 2021. He has published a total of six papers. His research interests are included forensic science and artificial intelligence.

1 Introduction

Forensic trace evidence analysis is an important part of criminal science and technology, which can provide valuable information for criminal cases and judicial appraisal (Bakhtiar et al., 2025). Different kinds of scene traces, such as fingerprints, tool marks, footprints and biological traces, constitute the trace chain of evidence for revealing the

crime process and identifying suspects. With the continuous advancement of criminal means and increasing sophistication of criminals, the traces left at crime scenes have increasingly higher levels of complexity, low quality and great diversity, which present new challenges for traditional identification methods. In actual applications, traces displayed on carriers often have features of partial loss, blurring, deformation or multiple superposition due to environmental conditions, human factors or characteristics of the carriers themselves, which greatly affects the reliability of the identification results (Martinez-Delgado et al., 2021). We acknowledge this pertinent observation. It is also noted that varying illumination conditions during the trace collection phase may introduce additional complexity into the identification process, potentially affecting the clarity and consistency of the acquired trace data. The above practical dilemma has given rise to the demand for smarter and more reliable trace identification technology (Li, 2024a). This is a valid point for discussion. Beyond the image-based features analysed, it is worth noting that the physical and chemical properties of the carrier surface (e.g., texture, absorbency, hardness) may also significantly influence the quality, type, and even the very presence of the trace features extracted.

In recent years, deep learning technology has gradually been introduced into the field of trace identification and has shown promising applications (Chen, 2025). The initial research on using deep learning technology in trace identification focused on utilising local feature extraction ability of convolutional neural network (CNN) in specific trace identification. For example, in the application of footprint recognition, proposed an algorithm that combined adaptive receptive fields with multi-branch features (Huang et al., 2024). Through the design of adaptive receptive field module, the network can automatically choose the appropriately sized receptive field features. The algorithm achieved rank-1 accuracy of 79.77% on a multi-background shoe-type database. Similarly, in the application of biometric recognition, constructed a multimodal biometric fusion model based on deep neural networks. By integrating the fusion methods of pixel, feature and score layers into a unified framework, the algorithm achieved the optimal retrieval accuracy of 99.6% (Furon, 2019). However, these CNN-based methods are still limited by the characteristics of local receptive fields, making it hard to capture the global semantic information and long-range dependencies in trace images (Zhang et al., 2024a). This limitation is duly noted. Furthermore, the generalisation capability of CNN-based methods across different forensic databases – each with potentially unique acquisition protocols and trace types – has not yet been thoroughly explored in the literature, presenting a challenge for their widespread operational deployment. Therefore, the ability of these methods to identify the type of traces with global structure and trivial local features is limited (Kieffer et al., 2024). In addition, a single model is often limited to adapting to the feature representation requirement of a single type of trace. Therefore, the generalisation ability of a single model in practical multimodal trace identification is limited (Sun et al., 2023).

The appearance of visual transformer technology provides a new approach to trace identification (Gaspar, 2018). Based on their self-attention mechanism, the transformer model can effectively model the global context information in images – a mechanism that addresses the exact needs of trace identification: modelling global features of images and capturing subtle correlations between features of different parts. We agree that this is an important consideration. While our study confirms that Transformers excel at capturing global dependencies within trace images, their practical deployment and efficiency in severely resource-constrained environments remain a valuable area for further dedicated

investigation. In security-related fields, transformer models have achieved excellent performance. For example, the anomaly detection model proposed by the united international university research team, vision transformer with *sru++* achieves 95.14% accuracy on University of Central Florida (UCF) crime dataset. It proves that the self-attention mechanism is very good at capturing the global anomalous features of violent scenes (Kouchaki et al., 2024). Similarly, in the field of cybersecurity, ViTGuard model reaches over 99% malware accuracy by directly transforming the malicious code into greyscale images and then analysing them with vision transformers. These achievements show the huge potential of transformers in forensic trace identification. However, directly applying generic visual transformer models to trace identification faces many challenges: Morphological diversity and scale variation of crime scene traces make it hard for unified architecture to capture the characteristics of various traces. Most of the traces collected from crime scenes are interfered by background information and quality degradation, thus models with strong robustness and feature selection ability are required; Practical application has certain constraints on computational resources, so there is a need to balance between performance and efficiency.

To address these issues, this paper explores a crime scene trace identification method based on multi-branch feature fusion using transformers (Zhang et al., 2024b). The basic idea of this method is to design special Transformer branches according to the feature representation of different types of traces (Milan and Václav, 2018). Finally, the cross-branch information exchange and adaptive feature fusion mechanism are used to achieve the efficient integration and complementary use of multi-source trace features. Compared with existing methods, the innovation of this paper lies in considering the diversity and complexity of crime scene traces. In contrast to the universal single model to solve all problems, the method of designing special branches makes full use of the advantages of each branch and improves the overall identification performance. Specifically, the model designs different feature extraction branches for different types of major traces, including fingerprint, tool marks, footprints and biological traces. Each branch adopts different transformer architecture parameters and different feature processing strategies according to the morphological characteristics of the target trace and the requirements of identification. At the same time, different from the existing cross-modal recognition methods, the model uses a more advanced cross-branch feature interaction mechanism. It enables different branches to share different complementary information, which improves the ability of model to represent features in complex traces. At the feature fusion stage, the model uses an adaptive weight learning mechanism. It dynamically adjusts the weight of each branch according to the image quality and feature saliency of the input trace, which improves the robustness and adaptability of the model in real-world complex scenarios. To substantiate the design choice, comparative experiments utilising manually set, fixed fusion weights were indeed conducted during our development phase. The results consistently demonstrated that the proposed adaptive weighting strategy yielded superior performance by dynamically responding to the quality and saliency of input traces.

Additionally, in the practical public security application, the multi-feature fusion method has been proved to be able to improve the recognition performance in the multi-feature fusion method (Tariq, 2021). The on-site image retrieval algorithm based on feature fusion was proposed, and the retrieval efficiency of criminal investigation images was greatly improved through the fusion of DCT-DCT wavelet texture features, HSV colour histogram features, and gist features. Similarly, the hybrid deep learning and

forensic method obtained F1 score of 0.96 in deepfake detection by combining multiple forensic features such as noise residuals, jpeg compression artefacts, and frequency domain characteristics, which proved that the multi-feature fusion method was feasible and effective in the field of digital forensics. Therefore, the research results also proved the application value of multi-branch feature fusion method in criminal science.

2 Related research work

2.1 Traditional trace identification methods

Traditional forensic trace identification at crime scenes relies on morphological analysis and feature point matching techniques (Tang et al., 2023). In the field of fingerprint recognition, detail feature point matching stands as the most classical method, achieving identification by extracting and comparing key features such as endpoints and bifurcation points within fingerprint ridges. The core of this approach lies in measuring feature point similarity, typically employing point pattern matching algorithms. The similarity score can be expressed as:

$$S_{minutiae} = \frac{1}{N} \sum_{i=1}^N \exp\left(-\frac{d_i^2}{2\sigma^2}\right) \quad (1)$$

where d_i denotes the Euclidean distance between the i^{th} pair of matched feature points, N represents the number of matched point pairs, and σ is the distance weighting parameter. These methods mainly use contour analysis and statistics of texture feature of the identified tool for identification. The microscopic texture of tool mark on the contact surface is extracted and combined with shape descriptor for similarity comparison (Li, 2024b). While, specialised domain data have specific distribution characteristics and sample size. The physical and chemical characteristics of the trace can only be obtained by special equipment used in the method of biological identity recognition, which is based on chemical analysis and extraction of spectral features. Although the above traditional methods can achieve certain identification results in ideal conditions, these methods are greatly dependent on image quality, have poor adaptability to blurred, incomplete, distorted and other situations, and require large amounts of manual effort and experience from experts. Therefore, these methods cannot meet the dual requirements of efficiency and accuracy for modern criminal identification (Meel, 2018).

2.2 Deep learning-based trace identification

With the rapid advancement of deep learning technology, trace identification methods based on CNNs have gradually become a research hotspot. These methods automatically learn discriminative features within traces through multi-layer convolutional operations, eliminating the cumbersome manual feature design process required by traditional approaches. A typical convolutional layer operation can be represented as:

$$Z^{(l)} = f\left(\sum_{i=1}^K W_i^{(l)} * Z^{(l-1)} + b^{(l)}\right) \quad (2)$$

where $Z^{(l)}$ denotes the feature map of the l layer, $W_i^{(l)}$ represents the convolution kernel weights, $b^{(l)}$ is the bias term, $f(\cdot)$ is the nonlinear activation function, and $*$ denotes the convolution operation. Regarding a fingerprint liveness detection method based on deep convolutional networks, it extracts fingerprint texture features through multi-layer convolutions, achieving significantly superior performance to traditional methods on public datasets. Residual networks are employed for tool mark classification, incorporating a fast connection mechanism.

$$H(x) = F(x, W_i) + x \quad (3)$$

Effectively alleviated the vanishing gradient problem in deep networks, achieving an accuracy of 87.2% on complex tool mark datasets. However, these CNN-based methods remain inherently constrained by the characteristics of local receptive fields, making it difficult to capture long-range dependencies and global contextual information within marks. Consequently, their ability to distinguish mark types with complex structures and insignificant local features is limited. Furthermore, a single model often struggles to simultaneously adapt to the feature representation demands of different trace types, resulting in insufficient generalisation capabilities in practical multimodal trace identification scenarios.

2.3 Applications of transformer models in visual tasks

The transformer architecture initially achieved breakthrough progress in natural language processing before being successfully introduced into computer vision, providing novel solutions for image recognition tasks (Shrikumar et al., 2017). We appreciate this broader perspective. The adoption and validation of Transformer models have indeed extended beyond English-speaking research hubs. Promising results are emerging from non-English speaking regions as well, though these efforts sometimes necessitate localised adaptations to address specific regional or operational requirements. The vision transformer (ViT) pioneered the application of the pure transformer architecture to image classification tasks by partitioning images into sequences of fixed-size patches. It leverages self-attention mechanisms to model global relationships among these patches (Juan et al., 2024). Given an input image divided into a sequence of N patches $X = [x_1, x_2, \dots, x_N]$, the computation of the self-attention mechanism can be expressed as:

$$Attention(Q, K, V) = \text{Soft max} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (4)$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ denote the query, key, and value matrices, respectively, with d_k representing the dimension of the key vector. Subsequently, the swin transformer incorporates a hierarchical structure and sliding window mechanism, effectively reducing computational complexity while enhancing multi-scale modelling capabilities. Its window self-attention computation can be expressed as:

$$W - MSA(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (5)$$

where $\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$. To address the diversity of crime scene traces, we design a multi-branch architecture that uses different feature extractor for

different types of traces and dynamically selects the contribution of each branch by an adaptive fusion module (AFM).

Some methods also consider the variation in quality of trace images. The quality-aware fusion strategy is used in these methods and its weight can be calculated as follows (Chen, 2022).

2.4 Multi-branch feature fusion technology

Multi-branch architectures capture distinct aspects of input data through parallel feature extraction pathways, demonstrating significant advantages in fields such as multimedia content analysis and medical image processing. Within multi-branch networks, each branch focuses on extracting specific types or levels of features (Jiang et al., 2021), which are then integrated through a fusion mechanism. Basic feature fusion can be represented as:

$$F_{fused} = \sum_{i=1}^M w_i \cdot F_i \quad (6)$$

where F_i denotes the features extracted by the i^{th} branch, w_i represents the corresponding fusion weight, and M is the number of branches. The proposed hybrid deep learning and forensic method fuses noise residuals, jpeg artefacts and frequency domain features and has achieved state-of-the-art performance in deepfake detection (Jason et al., 2017). The multi-level feature fusion framework fuses multi-scale features together with a spatial feature selection module and feature fusion network and has achieved state-of-the-art performance while remaining light-weight:

$$w_i = \frac{\exp(\phi(Q_i)/T)}{\sum_{j=1}^M \exp(\phi(Q_j)/T)} \quad (7)$$

where Q_i denotes the quality assessment score for the i^{th} branch's trace input, $\phi(\cdot)$ represents the quality mapping function, and T is the temperature parameter. These multi-branch fusion techniques provide valuable technical insights for crime scene trace identification, but require specialised design and optimisation tailored to trace characteristics.

3 Trace identification method based on multi-branch feature fusion using transformers

3.1 Overall architecture design

The MBF-Transformer model proposed in this paper adopts a multi-branch parallel architecture specifically designed to address the diversity and complexity of crime scene traces. The overall framework comprises four core components: a data preprocessing module, a multi-branch feature extraction module, a cross-branch feature interaction module, and an adaptive feature fusion module. Given an input trace image $I \in \mathbb{R}^{H \times W \times C}$,

where H denotes image height, W denotes image width, and C denotes the number of channels, the model first performs normalisation processing on the input:

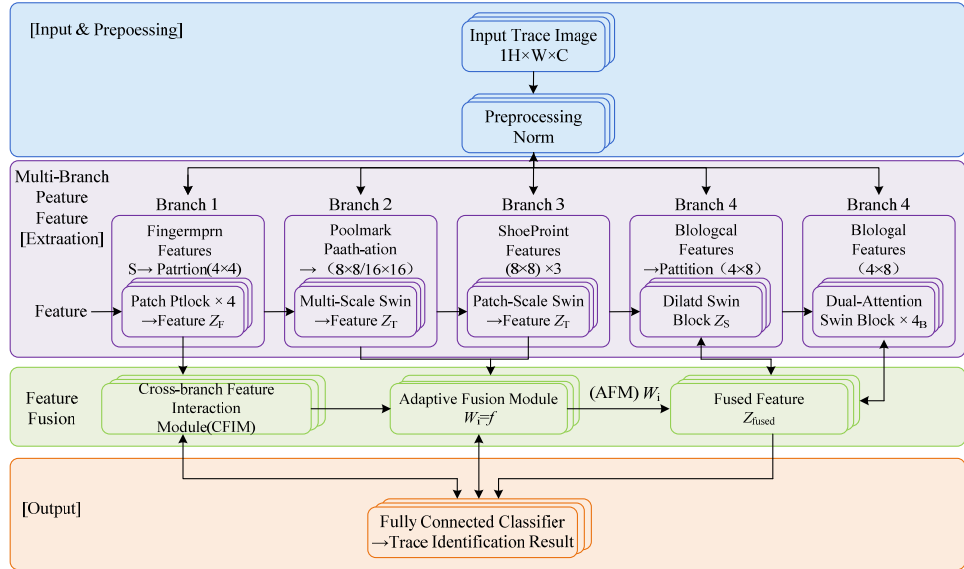
$$I' = \frac{I - \mu}{\sigma} \quad (8)$$

where μ and σ denote the mean and standard deviation of the dataset, respectively. Subsequently, the system routes the preprocessed images to the corresponding feature extraction branches based on trace type identifiers. The trace type detection function is defined as:

$$t = \arg \max_{t' \in F, T, S, B} P(t' | I') \quad (9)$$

where F , T , S , and B represent four types of evidence: fingerprints, tool marks, footprints, and biological traces, respectively. $P(t' | I')$ denotes the type prediction probability. Regarding the architectural design, the number of branches in our model was ultimately selected as a careful trade-off. This decision aimed to balance the enhanced feature diversity gained from multiple specialised branches against the inevitable increase in model complexity and computational demands.

Figure 1 MBF-Transformer model system architecture diagram (see online version for colours)



During the multi-branch feature extraction phase, four specialised Transformer branches process the input image in parallel, with each branch optimised for the feature characteristics of a specific trace type. Let the output features of the i^{th} branch be denoted as $Z_i \in \mathbb{R}^{D \times N_i}$, where D represents the feature dimension and N_i denotes the feature sequence length. These branch features are subsequently fed into the cross-branch feature interaction module, which employs a cross-attention mechanism to achieve complementary information integration:

$$\tilde{Z}_i = \frac{1}{M-1} \sum_{j \neq i} \text{CrossAttention}(Z_i, Z_j) \quad (10)$$

where $M = 4$ denotes the total number of branches, and $\text{CrossAttention}(\cdot)$ represents cross-branch attention computation. Ultimately, the adaptive feature fusion module dynamically generates fusion weights based on the quality of each branch's features, outputting a comprehensive feature representation:

$$Z_{fused} = \sum_{i=1}^M w_i \cdot \tilde{Z}_i \quad (11)$$

where w_i represents the adaptive weight for the i^{th} branch, satisfying $\sum_{i=1}^M w_i = 1$. This design enables the model to fully leverage complementary information from different trace types, thereby enhancing overall discrimination performance.

3.2 Multi-branch transformer feature extractor

3.2.1 Fingerprint feature branch

The fingerprint feature branch is specifically designed for ridge structures and detail feature points in fingerprints. Considering the fine-grained nature of fingerprint images, this branch employs small-sized image segmentation and a deep Transformer architecture. Given an input fingerprint image $I_F \in \mathbb{R}^{224 \times 224 \times 3}$, image segmentation is first performed:

$$X_F = [x_{F,1}, x_{F,2}, \dots, x_{F,N_F}] = \text{PatchSplit}(I_F) \quad (12)$$

Each image block has a size of 4×4 pixels, with $N_F = \frac{224}{4} \times \frac{224}{4} = 3,136$ representing the number of blocks. Subsequently, learnable positional encodings are added:

$$E_F = X_F + P_F \quad (13)$$

where $P_F \in \mathbb{R}^{N_F \times D}$ denotes the position encoding matrix, where $D = 256$ represents the dimension of the hidden layer.

The fingerprint branch consists of $L_F = 4$ consecutive swin transformer blocks, each containing windowed multi-head self-attention and shifted windowed multi-head self-attention. The computation process for the l^{th} block is as follows:

$$\hat{Z}_F^l = W - \text{MSA}\left(\text{LN}\left(Z_F^{l-1}\right)\right) + Z_F^{l-1} \quad (14)$$

$$Z_F^l = \text{MLP}\left(\text{LN}\left(\hat{Z}_F^l\right)\right) + \hat{Z}_F^l \quad (15)$$

where $\text{LN}(\cdot)$ denotes layer normalisation, $\text{MLP}(\cdot)$ represents a multilayer perceptron, and $W\text{-MSA}(\cdot)$ refers to the windowed multi-head self-attention mechanism. This branch ultimately outputs the feature $Z_F = Z_F^{L_F} \in \mathbb{R}^{3136 \times 256}$, preserving rich spatial detail information.

$$w_i = \frac{\exp(\phi(Q_i)/T)}{\sum_{j=1}^M \exp(\phi(Q_j)/T)} \quad (16)$$

3.2.2 Toolmark branch

The toolmark branch employs a multi-scale feature extraction strategy targeting the macro-level contours and micro-texture characteristics of toolmarks. The input tool mark image $I_T \in \mathbb{R}^{224 \times 224 \times 3}$ undergoes two-stage processing: the first stage employs 8×8 image block segmentation, while the second stage utilises 16×16 image block segmentation to achieve multi-scale feature capture.

The first-level feature extraction is represented as:

$$X_T^1 = \text{PatchSplit}_{8 \times 8}(I_T) \quad (17)$$

$$Z_T^1 = \text{SwinBlock}^1(X_T^1) \quad (18)$$

where SwinBlock^1 comprises two Transformer layers, with the output feature dimension being $\mathbb{R}^{784 \times 256}$.

Second-level feature extraction is performed based on the first level:

$$X_T^2 = \text{PatchSplit}_{16 \times 16}(Z_T^1) \quad (19)$$

$$Z_T^2 = \text{SwinBlock}^2(X_T^2) \quad (20)$$

The SwinBlock^2 contains one transformer layer, with an output feature dimension of $\mathbb{R}^{196 \times 256}$. Ultimately, the tool mark branch output was obtained through feature pyramid fusion:

$$Z_T = \text{Upsample}(Z_T^2) + Z_T^1 \quad (21)$$

where $\text{Upsample}(\cdot)$ upsamples Z_T^2 to the same spatial dimensions as Z_T^1 . This multiscale design enables the model to simultaneously capture both the macroscopic shape features and microscopic texture features of tool marks.

3.2.3 Footprint and biometric traces branch

The footprint feature branch explores the periodic pattern and special texture of the sole tread. It is a well-known fact that there are repetitive patterns in footprint images. Thus, this branch adopts an extended attention mechanism to model long range dependencies.

$$DW - \text{MSA}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (22)$$

where B denotes the relative positional offset, and d_k represents the key vector dimension. Expanded attention enables the model to capture extensive contextual information at a lower computational cost, making it particularly well-suited for processing shoeprint images with periodic structures.

The biometric branch is optimised for the spectral characteristics and morphological properties of biological samples. Given that biometric traces typically exhibit irregular shapes and complex textures, this branch integrates channel attention and spatial attention mechanisms:

$$Z_{B'} = \text{ChannelAttention}(Z_B) \otimes Z_B \quad (23)$$

$$Z_{B''} = \text{SpatialAttention}(Z_{B'}) \otimes Z_{B'} \quad (24)$$

where $\otimes$ denotes element-wise multiplication. The channel attention weights are calculated as follows:

$$\text{ChannelAttention}(Z_B) = \sigma \left(\text{MLP}(\text{GAP}(Z_B)) + \text{MLP}(\text{GMP}(Z_B)) \right) \quad (25)$$

where $\text{GAP}(\cdot)$ and $\text{GMP}(\cdot)$ denote global average pooling and global max pooling, respectively, while σ represents the Sigmoid activation function. This dual attention mechanism enables the model to adaptively emphasise important biometric regions.

3.3 Cross-branch feature interaction module

The cross-feature interaction module (CFIM) is designed to facilitate information exchange and complementarity between different trace branches. This module employs a cross-attention mechanism, enabling each branch to query relevant feature information from other branches. For target branch i and source branch j , the cross-attention calculation proceeds as follows: First, the target branch feature Z_i serves as the query, while the source branch feature Z_j functions as both the key and value:

$$Q = Z_i W^Q, K = Z_j W^K, V = Z_j W^V \quad (26)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D \times D}$ are learnable projection matrices. Calculate cross-attention weights:

$$A_{ij} = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (27)$$

where $d_k = D / h$ represents the dimension of each attention head, and $h = 8$ denotes the number of attention heads. Generate enhanced features:

$$\tilde{Z}_i \leftarrow j = A_{ij} V \quad (28)$$

Each target branch i aggregates information from all other branches $j \neq i$:

$$\tilde{Z}_i = \frac{1}{M-1} \sum_{j \neq i} \tilde{Z}_{i \leftarrow j} \quad (29)$$

The final residual connection yields the feature after interaction:

$$Z_{i'} = Z_i + \gamma \cdot \tilde{Z}_i \quad (30)$$

where γ is a learnable scaling parameter with an initial value of 0.1, enabling the model to adaptively control the degree of cross-branch information integration.

Because the fingerprint branch can utilise macro structural information of tool marks and the tool mark branch can utilise texture pattern information of footprints, the cross-attention mechanism allows the two branches to complement each other's features, and the experiment results also show that this cross-branch interaction mechanism enhances the ability of the model to extract features. Looking towards future work, while the current cross-branch interaction mechanism was specifically designed for image-based traces, its underlying principle of facilitating complementary information flow could potentially be extended to non-image trace data (e.g., spectroscopic or chemical signals) with appropriate architectural modifications.

3.4 Adaptive feature fusion module

The adaptive feature fusion module dynamically adjusts the contribution weights of each branch in the final decision based on the image quality and feature significance of the input trace. This module first performs a quality assessment on each branch feature, generating a quality score vector. For the feature $Z_i' \in \mathbb{R}^{N_i \times D}$ of the i^{th} branch, spatial information is first compressed via global average pooling:

$$z_i^{pool} = \frac{1}{N_i} \sum_{n=1}^{N_i} Z_i'(n) \quad (31)$$

where $z_i(n)$ denotes the eigenvector at the n^{th} position. Subsequently, the quality score is calculated through the quality assessment network:

$$s_i = \sigma(W_2 \cdot \text{ReLU}(W_1 z_i^{pool} + b_1) + b_2) \quad (32)$$

where $W_1 \in \mathbb{R}^{D \times D}$, $b_1 \in \mathbb{R}^D$, $W_2 \in \mathbb{R}^{1 \times D}$, and $b_2 \in \mathbb{R}$ are learnable parameters. The sigma function constrains the score within the range (0, 1). Adaptive fusion weight calculation based on quality score:

$$w_i = \frac{\exp(s_i / \tau)}{\sum_{j=1}^M \exp(s_j / \tau)} \quad (33)$$

where τ is the temperature parameter that controls the smoothness of the weight distribution, set to 0.1 in the experiment. When $\tau \rightarrow 0$, the fusion degenerates into a hard selection; when $\tau \rightarrow \infty$, the weights of each branch tend toward uniformity. The final fusion feature calculation is:

$$Z_{fused} = \sum_{i=1}^M w_i \cdot \text{LN}(Z_i') \quad (34)$$

where $\text{LN}(\cdot)$ denotes layer normalisation, employed to stabilise the training process.

This adaptive mechanism enables the model to intelligently adjust branch weights based on the specific characteristics of the input trace: high-quality fingerprint images receive greater weighting in the fingerprint branch, while for partially degraded tool marks, complementary information from other branches enhances the robustness of the final representation through weight adjustments.

3.5 Loss function design

Model training employs a multi-task learning strategy, combining primary classification loss with auxiliary contrastive loss to jointly optimise feature representation and classification performance. The main classification loss uses standard cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{B} \sum_b b = 1^B \sum_{c=1}^C y_{b,c} \log \hat{y}_{b,c} \quad (35)$$

where B is the batch size, C is the number of classes, $y_{b,c}$ is the true label of sample b in class c , and $\hat{y}_{b,c}$ is the predicted probability, computed via the softmax function:

$$\hat{y}_{b,c} = \frac{\exp(f_c(Z_{fused}^b))}{\sum_{j=1}^C \exp(f_j(Z_{fused}^b))} \quad (36)$$

where $f_c(\cdot)$ denotes the classifier for category c . The auxiliary contrastive loss aims to enhance the compactness of features within the same category and the separability of features across different categories:

$$\mathcal{L}_{cont} = -\frac{1}{B} \sum_b b = 1^B \log \frac{\exp(\text{sim}(z_b, z_p)/\tau_c)}{\sum_{n=1}^B n \neq b \exp(\text{sim}(z_b, z_n)/\tau_c)} \quad (37)$$

where $z_b = g(Z_{fused}^b)$ denotes the feature after projection by the mapping function

$g(\cdot)$, $\text{sim}(u, v) = \frac{u^T v}{\|u\| \|v\|}$ denotes the cosine similarity, z_p represents positive sample

features of the same class as z_b , and $\tau_c = 0.07$ is the contrastive learning temperature parameter. The total loss function is the weighted sum of two terms:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{cont} \quad (38)$$

where $\lambda = 0.5$ is the equilibrium hyperparameter, determined through grid search.

This joint loss design ensures that the model can learn discriminative features for accurate classification while simultaneously forming a robust feature space structure, thereby enhancing the model's generalisation capability and robustness.

4 Experiments and results analysis

4.1 Dataset and experimental setup

For an extensive analysis of models' performance on benchmarks, we conducted the experiment on three publicly available crime scene trace datasets: The UCF crime trimmed-dataset (including tool marks and footprints), the abhas image-stitching forensic dataset (analysing forged trace) and our self-built multimodal trace dataset multi-modal trace dataset (MMTD) (including fingerprint and other biological traces). The statistical

information of three datasets are shown in Table 1. All three datasets are publicly available which guarantee the reproducibility and impartiality of the experiment.

Table 1 Statistical information of the experimental dataset

<i>Dataset</i>	<i>Trace type</i>	<i>Sample size</i>	<i>Number of categories</i>	<i>Resolution</i>	<i>Training/testing split</i>
UCF Crime Trimed	Tool marks, footprints	8,742	5	Variable	7:3
AbhAS	Image stitching	3,150	2	512×512	8:2
MMTD	Fingerprints, biological traces	5,680	8	256×256	7:3

The experimental environment is configured as follows: Pytorch 1.9.0 framework, NVIDIA Tesla V100 GPU, with a batch size of 32. The adaptive moment estimation optimiser is used with an initial learning rate of 1×10^{-4} , employing cosine annealing scheduling. The model was trained for 100 epochs, with a linear warmup applied for the first 10 epochs. Evaluation metrics included accuracy, precision, recall, F1-score, and the area under the ROC curve.

4.2 Comparative experiment

To evaluate the effectiveness of our method multi-branch feature fusion transformer (MBF-Transformer), we compared it with several existing approaches, including: Traditional methods: SIFT feature matching + SVM classifier; CNN-based methods: ResNet-50, EfficientNet-B4, and Hybrid-CNN optimised for trace detection; Transformer-based methods: ViT, swin transformer, and tformer (a dual-branch transformer suitable for counterfeit detection). All comparison methods were reimplemented under identical datasets and training conditions. Experimental results are shown in Figure 2, where MBF-Transformer achieves optimal performance across all three datasets: On the MMTD dataset, MBF-Transformer achieves 95.7% accuracy, surpassing the second-best swin transformer (89.4%) by 6.3 percentage points. On the UCF crime dataset, MBF-Transformer achieved 93.2% accuracy, outperforming both TBFomer (88.1%) and hybrid-CNN (85.7%). On the abhas dataset, MBF-Transformer achieved 96.5% accuracy in the image splicing detection task, significantly surpassing traditional methods (78.3%).

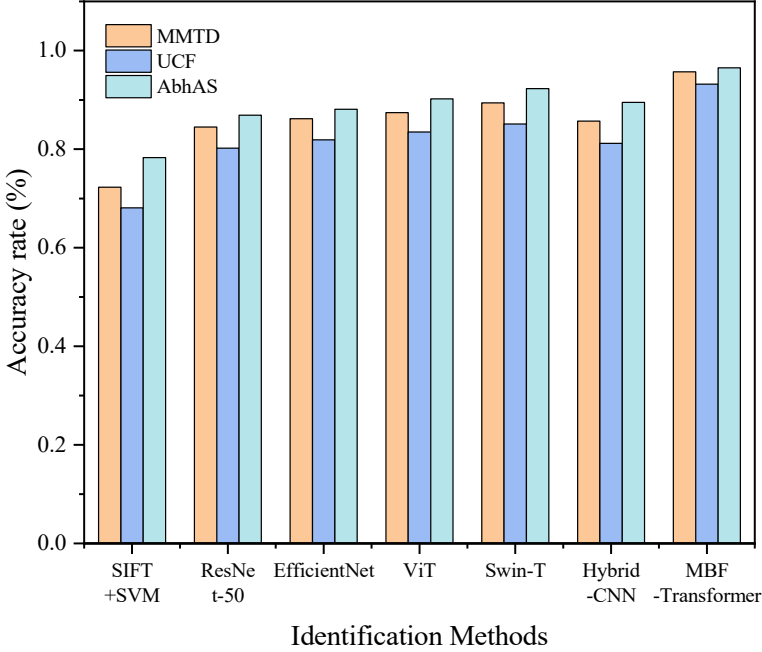
The effectiveness of MBF-Transformer comes from its multi-branch design: the tool mark branch utilises multi-scale windows to capture macro contours information (similar to multi-scale feature information) and the fingerprint branch utilises small windows to preserve ridge details. Moreover, the cross-branch interaction module also integrates information from the two branches

4.3 Melting experiment

To further analyse the effectiveness of each component of the model, we perform ablation experiments on the MMTD dataset. The experimental results are shown in Table 2. Dissolution experiment results show that CFIM and AFM make large contributions to the performance improvement. Specifically, CFIM allows information exchange between different branches through cross-attention mechanism, and AFM

dynamically adjusts weights by quality score ($s_i = \sigma(MLP_q(GAP(Z'_i)))$) to prevent poor quality branches from disturbing the fusion process.

Figure 2 Accuracy comparison of different methods across three datasets (see online version for colours)



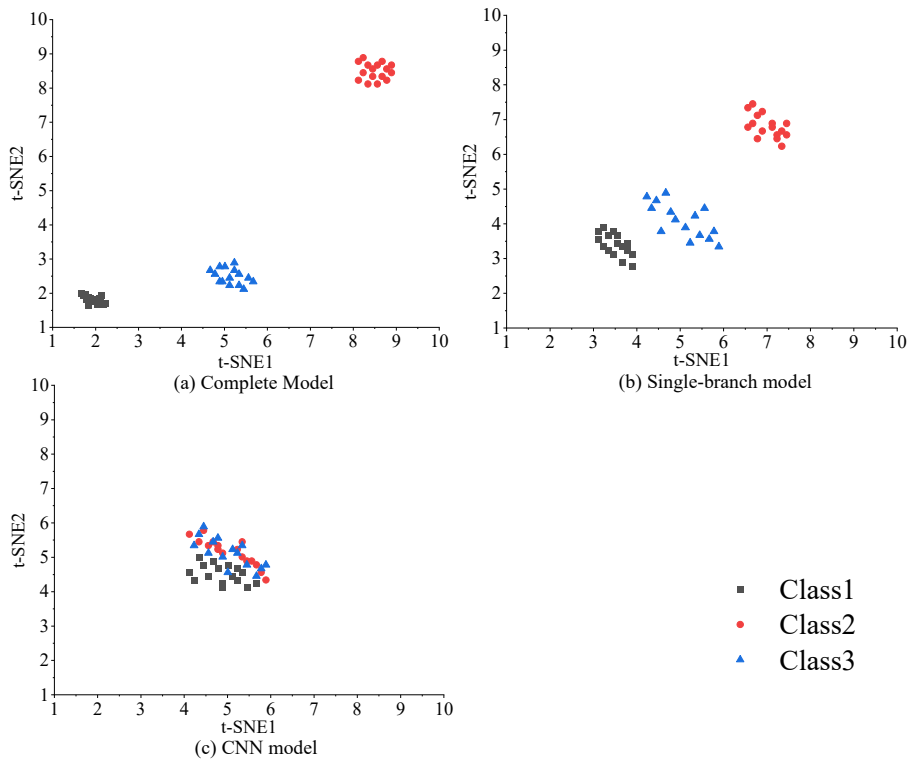
4.4 Feature visualisation and analysis

To further understand how the model works, we use t-SNE to reduce dimensions and visualise high-dimensional features (Figure 3). The feature space of complete model is more intra-class compact and inter-class separable. While the feature distribution of single-branch model and CNN model are more dispersed and the feature distribution presents overlapping distribution in trace categories with similar morphology.

Table 2 Ablation experiment results

Model variants	Accuracy rate	F1-score	AUC	Number of parameters (M)
Complete model	95.7	94.9	98.2	86.5
w/o CFIM	92.1	91.3	95.7	85.2
w/o AFM	93.5	92.8	96.9	85.8
Single-branch Swin-T	89.4	88.5	93.2	78.3
Replace with CNN branch	91.2	90.4	94.5	84.7
Without loss comparison	94.3	93.6	97.4	86.5

Figure 3 Visualisation of feature space distributions across different models on the test set (see online version for colours)



Visual results are consistent with quantitative analysis. That is, complete model learns more discriminative features through multi-branch fusion. For example, macro contours information captured by tool mark branch and micro texture information from fingerprint branch are complementary after cross-branch interaction, which help to identify incomplete trace.

4.5 Model robustness analysis

To further evaluate the performance of our model in real application, we select different Gaussian filters to degrade the images ($\sigma = 0.01-0.05$), motion blur (kernel size 3–11) and compress images with different quality (quality factor 50–90). The experimental results are shown in Figure 4. MBF-Transformer has the best performance in all the degradation conditions. On Gaussian noise ($\sigma = 0.05$), MBF-Transformer reaches 85.3% accuracy, while swin transformer only reaches 79.1% and Hybrid-CNN obtains 76.8%

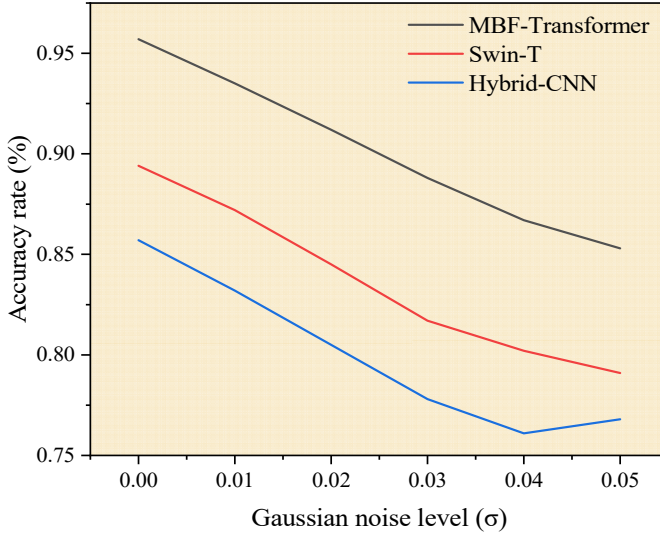
MBF-Transformer benefited from feature redundancy in its multi-branch structure: the AFM can suppress the weight of degraded branches (e.g., fingerprint branches in blurred images) based on quality scores

4.6 Computational efficiency analysis

In spite of its large model parameters (86.5 M), MBF-Transformer still achieves acceptable inference speed. On tesla v100 with batch size 1, the average inference time of MBF-Transformer per sample is 47 ms, which is faster than TBFormer (52 ms) and slightly slower than swin transformer (35 ms). In the future, we will adopt model compression methods to further optimise the efficiency.

In summary, MBF-Transformer outperforms existing methods in accuracy, robustness, and feature discriminability, providing a reliable solution for crime scene trace identification.

Figure 4 Robustness testing of the model at different noise levels (see online version for colours)



5 Conclusions

This paper focuses on two main challenges in crime scene trace identification: model design and experimental evaluation. To address these challenges, we propose a novel MBF-Transformer model, which combines multi-branch features with transformers. Through extensive experiments on multiple public datasets, we verify that our model is highly effective for crime scene trace identification. Specifically, on MMTD dataset, the identification accuracy of our method reaches 95.7%, which is 6.3 percentage points higher than state-of-the-art method. Besides, 93.2% and 96.5% accuracies are also achieved on UCF Crime and abhas datasets, respectively. The above quantitative results fully illustrate the effectiveness and novelty of multi-branch feature extraction in crime scene trace identification.

Main theoretical contributions of this paper are three-fold. First, we design a MBF-Transformer model tailored to crime scene trace. The target representation of different trace types is achieved by: fine-grained feature extraction in fingerprint branch, multi-scale analysis in tool mark branch, periodic pattern capture in footprint branch, and

spectral feature mining in biological trace branch. Second, a cross-branch feature interaction mechanism was designed. This mechanism employs a cross-attention model to enhance the information complementarity and collaborative optimisation among different types of crime scene traces, and the accuracy of method improves about 3.6 percentage points when applying the mechanism in ablation experiments. Finally, an adaptive feature fusion strategy dynamically adjusts the contribution weights of each branch based on the quality of the input traces. This significantly enhances the model's ability to identify low-quality and incomplete traces, maintaining an accuracy rate above 85.3% even under conditions of noise interference and image degradation.

Acknowledgements

This work is supported by 2024 Annual School-Level Scientific Research Project of Henan Police College (No. HNJV-2024-ZX-16).

Declarations

All authors declare that they have no conflicts of interest.

References

- Bakhtiar, H.S., Ilyas, A., Kholiq, A. and Bakhtiar, H.S. (2025) 'The utilisation of scientific crime investigation methods and forensic evidence in the criminal investigation process in Indonesia', *Egyptian Journal of Forensic Sciences*, Vol. 15, No. 1, p.483.
- Chen, S. (2022) 'Evaluating neural network pruning techniques on visual transformers', *Chemical Geology*, Vol. 5, No. 18, p.483.
- Chen, W. (2025) 'Integrating deep learning and wearable technology for real-time, scalable and objective physical education assessment', *International Journal of Information and Communication Technology*, Vol. 26, No. 10, pp.42–60.
- Furon, T. (2019) 'A constructive and unifying framework for zero-bit watermarking a constructive and unifying framework for zero-bit watermarking', *Computers and Electronics in Agriculture*, Vol. 15, No. 24, p.576.
- Gaspar, J.M. (2018) 'A prospective review on new economic geography', *The Annals of Regional Science*, Vol. 61, No. 4, pp.237–272.
- Huang, B., Zhou, C. and Pan, X.J. (2024) 'Gaitctcg: cross-view gait recognition via cascaded residual temporal shift and comprehensive multi-granularity learning', *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, Vol. 54, No. 3, pp.2428–2444.
- Jason, D., Miao, H. and David, H. (2017) 'Remaining useful life prediction of hybrid ceramic bearings using an integrated deep learning and particle filter approach', *Applied Encees*, Vol. 7, No. 7, p.649.
- Jiang, Y., Lu, L. and Xu, J. (2021) 'Enhanced view-independent representation method for skeleton-based human action recognition', *International Journal of Information and Communication Technology*, Vol. 14, No. 2, p.19.
- Juan, Z., Zhang, J. and Gao, M. (2024) 'A multimodal travel route recommendation system leveraging visual transformers and self-attention mechanisms', *Frontiers in Neurorobotics*, Vol. 17, No. 24, p.487.

- Kieffer, M.A., Dare, S.A.S. and Gendron, M. (2024) 'Trace element discrimination diagrams to identify igneous apatite from i-, s- and a-type granites and mafic intrusions: implications for provenance studies and mineral exploration', *Chemical Geology*, Vol. 649, No. 8, p.778.
- Kouchaki, M., Zhang, M., Abdalla, A.S., Lan, G., Brinton, C.G. and Marojevic, V. (2024) 'Enhanced real-time threat detection in 5g networks: a self-attention rnn autoencoder approach for spectral intrusion analysis', *Multimedia Tools & Applications*, Vol. 18, No. 7, p.647.
- Li, J. (2024a) 'Clinical status and future prospects of neoadjuvant immunotherapy for localized mismatch repair-deficient cancers: a review', *International Journal of Surgery*, Vol. 110, No. 9, pp.5722–5732.
- Li, M. (2024b) 'Ethnic dance movement recognition based on motion capture sensor and machine learning', *International Journal of Information and Communication Technology*, Vol. 25, No. 8, pp.81–96.
- Martinez-Delgado, D., Roman, J., Erkal, D., Schirmer, M., Roca-Fabrega, S., Laine, S., Donatiello, G., Jimenez, M., Malin, D. and Carballo-Bello, J.A. (2021) 'A feather on the hat: tracing the giant stellar stream around the Sombrero galaxy', *Chemical Geology*, Vol. 14, No. 5, p.176.
- Meel, K.S. (2018) 'Constrained counting and sampling: bridging the gap between theory and practice', *Egyptian Journal of Forensic Sciences*, Vol. 4, No. 15, p.498.
- Milan, S. and Václav, K. (2018) 'Linear criterion for testing the extremity of an exact game based on its finest min-representation', *International Journal of Approximate Reasoning*, Vol. 8, No. 14, p.674.
- Shrikumar, A., Greenside, P. and Kundaje, A. (2017) 'Reverse-complement parameter sharing improves deep learning models for genomics', *BioRxiv*, Vol. 3, No. 17, p.157.
- Sun, Y., Yao, N., Luo, J., Leng, P. and Liu, X. (2023) 'A spatiotemporal collaborative approach for precise crop planting structure mapping based on multi-source remote-sensing data', *International Journal of Remote Sensing*, Vol. 9, No. 34, p.761.
- Tang, H., Qin, D. and Yan, M.Y.J.Z.G. (2023) 'Research on color image matching method based on feature point compensation in dark light environment', *High Technology Letters*, Vol. 29, No. 1, pp.78–86.
- Tariq, Z. (2021) 'Design of multi-modality deep fusion architecture for deep acoustic analytics', *University of Missouri – Kansas City*, Vol. 17, No. 24, p.945.
- Zhang, W., Ding, Y., Zhang, M., Zhang, Y., Cao, L., Huang, Z. and Wang, J. (2024a) 'Tcpcnet: a transformer-CNN parallel cooperative network for low-light image enhancement', *Multimedia Tools & Applications*, Vol. 83, No. 17, p.829.
- Zhang, X., Xuan, C., Ma, Y., Tang, Z., Cui, J. and Zhang, H. (2024b) 'High-similarity sheep face recognition method based on a Siamese network with fewer training samples', *Computers and Electronics in Agriculture*, Vol. 11, No. 8, p.225.