npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01831-7

Hyperspectral imaging based multidimensional Terracotta Warrior extraction algorithm

Check for updates

Shi Qiu, Pengchang Zhang ✉, Siyuan Li & Bingliang Hu

The Terracotta Warriors of the Qin Shi Huang Mausoleum hold immense historical significance. However, during the excavation process, accurately locating and extracting the Terracotta Warriors presents substantial challenges. We conducted research to develop an extraction algorithm. (1) A hyperspectral model was established to extract features. (2) A visual and distance-based feature extraction model was established to simulate visual and distance information, enabling the extraction of multidimensional features. (3) A parallel deep network was constructed, integrating spectral information, simulated visual data, and distance information to accurately extract the Terracotta Warriors. To verify the effectiveness of the algorithm, we collected 51 sets of large-size hyperspectral data on site. Experimental results demonstrate that the algorithm achieves a Combination Measure of 93%, outperforming the traditional Graph-FCN algorithm with a score of 12%. The algorithm innovatively incorporates distance information to classify the Terracotta Warriors, achieving good results.

The Qin Dynasty was the first unified feudal dynasty in Chinese history. The Terracotta Warriors of the Qin Shi Huang Mausoleum carry the profound history and rich culture of the Chinese nation, spanning thousands of years[1]. Research on these relics holds significant cultural importance. However, during the excavation of the Terracotta Warriors, accurately locating their position, environment, and depth remains a key challenge and is currently a hot research topic.

Current research on information acquisition methods mainly includes: (1) Based on 3D point cloud: Cheng et al.[2] used point cloud data to analyze the corner data of the Terracotta Warriors. Hu et al.[3] used 3D scanning to obtain facial information of the Terracotta Warriors and carry out facial feature analysis. Liu et al.[4] constructed AMS-Net to realize fragment analysis of the Terracotta Warriors. Based on the obtained point cloud data, Hu et al.[5] established EGG-Net to achieve self-supervised segmentation of the Terracotta Warriors. Lei et al.[6] analyzed the structure of the Terracotta Warriors based on 3D scanning model. Gao et al.[7] established the Fractional Fourier Transform to realize 3D imaging of the Terracotta Warriors. (2) Based on an RGB image: Sheng et al.[8] adopted a deep-learning algorithm to detect and recognize the face of the Terracotta Warriors. Yang et al.[9] classify the Terracotta Warriors fragments based on spatial and texture information. Li et al.[10] realized virtual restoration of cultural relics based on the Fuzzy Logic Algorithm. Rui et al.[11] constructed a Residual Network to spline the Terracotta Warriors. (3) Based on microscopic images: Wang et al.[12] obtained the microbial information of the Terracotta Warriors and analyzed it. Wang et al.[13] assisted the restoration of the Terracotta Warriors with microscopic imaging technology. (4) Based on non-imaging spectral information: Cui et al.[14] analyzed the hidden information of cultural relics using terahertz. Gou et al.[15] analyzed the composition of the Terracotta Warriors from a materials perspective and realized the restoration. Xia et al.[16] analyzed the environment of the Terracotta Warriors and gave suggestions on their preservation. Zhang et al.[17] studied the corrosion degree of the Terracotta Warriors containing bronze weapons. With the development of imaging technology, a new imaging technology came into being. The most typical is hyperspectrum, which has spatial and spectral resolution, and its spectral information can reflect the material information of ground objects, which has a very large application prospect[18]. The characteristics of different acquisition methods are shown in Table 1.

The research of hyperspectral in cultural relics mainly includes: (1) Algorithms based on traditional features: Zeng et al.[19] reviewed the application of spectral technology in cultural relics analysis, and explained the feasibility of hyperspectral application in cultural relics. Han et al.[20] used hyperspectral technology to analyze the color of the surface of cultural relics. Feng et al.[21] identified the authenticity of porcelain through spectral technology. Wei et al.[22] identified the color of cultural relics based on hyperspectrum. (2) Algorithm based on deep learning: Qiu et al.[23] used hyperspectral technology to excavate Sanxingdui cultural relics information.

Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, P. R. China.
✉e-mail: zhangpengchang@opt.ac.cn

Table 1 | Analysis of information acquisition methods for the Terracotta Warriors

Method	Advantages	Disadvantages
3D	Complete 3D spatial information.	Large data volume, only structural information is represented, and unable to conduct further in-depth research.
RGB	Acquires 2D spatial information and color information, conforms to human visual perception.	Unable to conduct further in-depth research from a material perspective.
Microscopy	Can obtain microscopic information, indicating subtle changes.	Unable to obtain more macroscopic analysis.
Non-imaging spectroscopy	Can represent information at different spectral segments, reflecting material information to a certain extent.	Can only represent information over a smaller area.
Imaging spectroscopy	Has spatial and spectral information, capable of analysis from both macro and micro perspectives.	Large data volume, the extraction of typical features is relatively difficult.

Table 2 | Analysis of algorithms based on hyperspectral imaging research

Method	Advantages	Disadvantages
Based on traditional features	Features have strong representability, and a large part of the features can be quantified.	Limited feature generalization ability, and the scope of features is relatively narrow.
Based on deep-learning features	Simulates human cognitive processes, with strong feature generalization ability.	Features are more abstract, and interpretability is relatively weak.

Peng et al.[24] realized mural analysis by fusing hyperspectral image information. Liu et al.[25] analyzed the color of cultural relics based on the visible spectral segment. Wang et al.[26] constructed 3D image virtual restoration based on hyperspectral images. Pan et al.[27] extracted weak mural information based on spectral information. Peng et al.[28] constructed a deep-learning framework to study the painting background. Therefore, the characteristics of the algorithm research based on hyperspectral imaging can be summarized as shown in Table 2.

In summary, research on the feature analysis of terracotta figurines, represented by the Terracotta Warriors, is of great significance. The main challenges faced are: (1) Insufficient characterization using a single sensor. (2) Difficulty in aligning data from multiple sensors. (3) Difficulty in effective feature selection.

As discussed above, hyperspectral imaging has become an effective tool for relic analysis due to its spatial and spectral features. Therefore, (1) Hyperspectral imaging is used to characterize features from both spatial and spectral dimensions. (2) Distance information is simulated based on specific spectral bands to avoid sensor alignment issues. (3) Two collaborative deep extraction networks based on deep-learning algorithms are constructed to achieve accurate extraction of the Terracotta Warriors.

Methods

Hyperspectral imagery contains information from multiple spectral bands, each representing distinct characteristics. To fully utilize the unique features of each spectral band, (1) the hyperspectral image sequence is used as input to establish a feature extraction algorithm based on hyperspectral data. (2) Relevant spectral band information is extracted to simulate the visual perception process, generating synthesized RGB images to extract visual information of the target. Based on the correspondence between specific spectral bands and distance, spectral information is extracted to represent distance information. A feature extraction algorithm based on visual and distance data is then constructed. (3) The results from the hyperspectral-based feature extraction algorithm and the visual-and-distance-based feature extraction algorithm are fused at the decision level to achieve the extraction of the Terracotta Warriors. Its process is shown in Fig. 1.

Hyperspectral-based feature extraction model

The hyperspectral-based feature extraction model consists of a data preprocessing module, a RSFEM module, a NA module, and a SSFFM module.

1. Hyperspectral imagery (HIS) contains information from multiple spectral bands. To represent hyperspectral features with fewer data, Principal Component Analysis (PCA) is employed for dimensionality reduction. PCA achieves this by transforming the underlying correlated variables into a smaller set of uncorrelated variables, known as principal components. Let HIS be denoted as X_{in}, with a resolution of $H\times W\times S$. The dimensionality of X_{in} is reduced through PCA as follows:

$$X_{PCA} = F(Xin) \tag{1}$$

where F represents the PCA operation, which retains the principal component information within the limited channels. The resolution of X_{PCA} is $H\times W\times B$.

2. Residual Spectral Feature Extraction (RSFEM) Module: To further extract useful shallow and deep features while reducing noise interference, the RSFEM module is constructed. This module employs multiple 3D convolution blocks to extract hyperspectral information. RSFEM consists of five sets of 3D modules, each of which includes convolutional layers, batch normalization layers (BatchNorm, BN), and activation layers using the ReLU function in series. The module takes X_{PCA} as input and applies a 3D convolution block with a kernel size of (1, 1, 5) to extract local spectral features. Since spectral mixing commonly occurs between adjacent pixels, two 3D convolution blocks with kernel sizes of (3, 3, 1) are applied to reduce redundancy caused by spectral mixing. Finally, the data size is adjusted through reconstruction. Three RSFEM modules are concatenated to extract features, and the output is denoted as X_R. To mitigate the problem of network degradation caused by excessively deep layers, residual modules are established between different convolution blocks. These residual modules fuse shallow features with subsequent deep features, enriching the spectral semantic information.
3. Neighboring Attention (NA) Module: Based on spatial correlation, the correlation between objects in an image is inversely proportional to their distance. In other words, the class of a central pixel is usually the same as the class of its neighboring pixels. Building on the RSFEM, the NA module is constructed to perform the calculation. The matrix **D** is generated by computing the Euclidean distance between the axial pixel points and the central pixel point:

$$D_{ij} = \sqrt{\sum_{k}^{B} (X_{ij}^{k} - X_{c}^{k})^{2}} \tag{2}$$

where D_{ij} is the Euclidean distance between the pixel in the ith row and jth column of the spatial block and the central pixel. B represents the number of

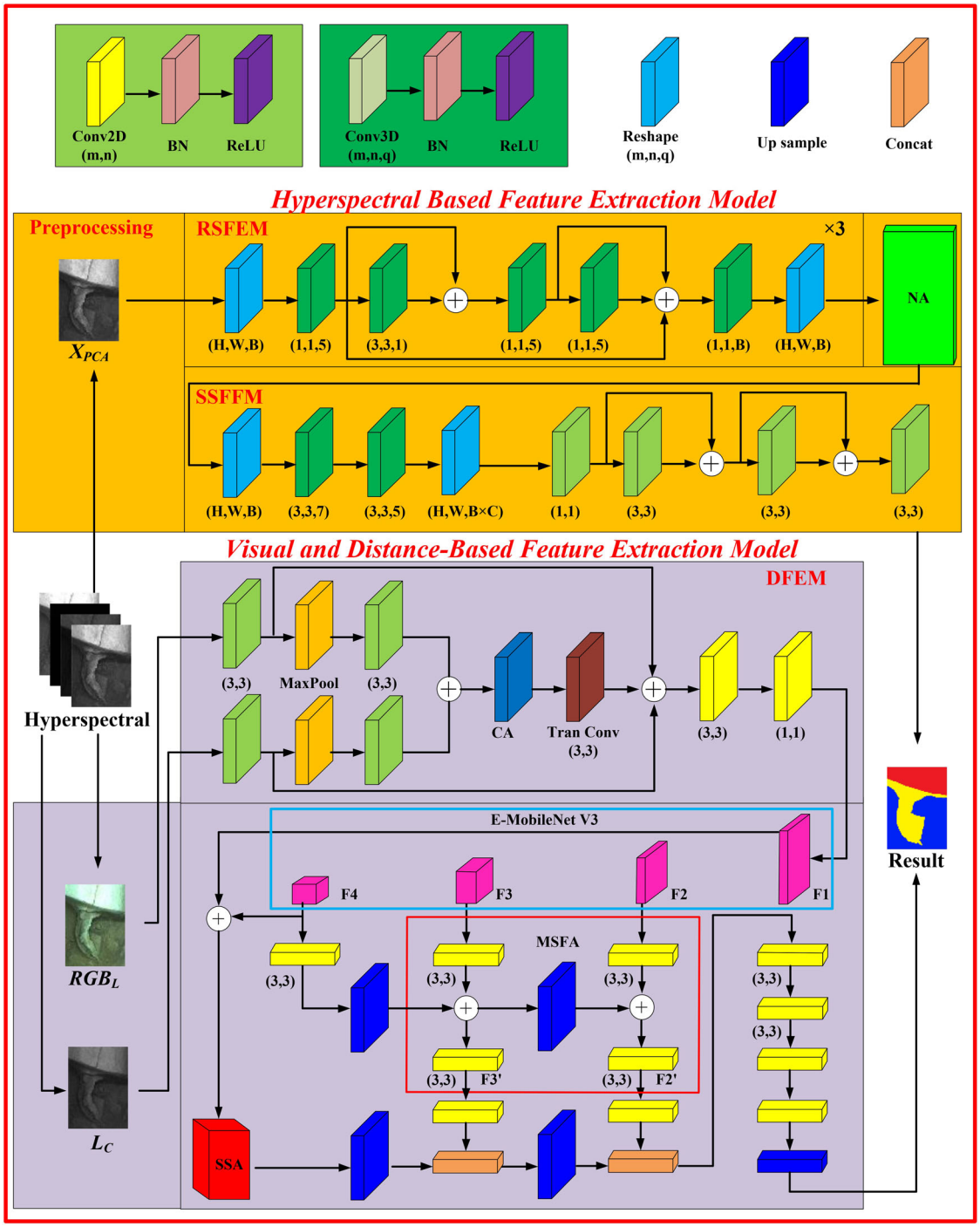


Fig. 1 | Algorithm flowchart.

spectral dimensions. X_{ij}^k is the feature value of the kth spectral dimension for the pixel in the ith row and jth column. X_c^k is the feature value of the kth spectral dimension for the central pixel of the spatial block.

With X_R as input, after weight assignment and residual addition, the final output of the module is:

$$X_{Ro} = X_R W + X_R \tag{3}$$

$$W_{ij} = 1 - \frac{D_{ij}}{Max(D_{ij})} \tag{4}$$

where W_{ij} is the weight assigned to the pixel in the ith row and jth column.

4. Spatial–Spectral Feature Fusion Module (SSFFM): Hyperspectral images contain both spectral and spatial information. The RSFEM module extracts spectral features from HIS, while the NA module extracts spatial features. To effectively fuse these spatial and spectral features, the SSFFM model is constructed, consisting of two 3D convolution blocks and four 2D convolution blocks, with pointwise convolution and residual connections introduced to further extract features. Let the input data for the module be X_{Ro}. Two 3D convolution blocks with kernel sizes of (3, 3, 7) and (3, 3, 5) are used to extract spatial–spectral information. The number of feature map blocks is gradually increased, and the size of the output feature map blocks is expanded. Next, the spectral and feature dimensions are merged to

obtain the feature map. A pointwise convolution block of size (1, 1) is used for dimensionality reduction, which reduces the computational load and also minimizes network parameters. Three additional (3, 3) convolution blocks are applied to obtain the final features. Residual connections are used between each layer to reduce training difficulty, prevent network degradation, and better integrate spatial–spectral information, enhancing the feature representation capability.

Visual and distance-based feature extraction model

Vision is the primary way of perceiving the existence of objects, with color and distance being its main representations. Therefore, three spectral bands that are closest to RGB from hyperspectral images are selected to form a pseudo-color image, RGB_L. In addition, during the hyperspectral data acquisition process, the light source is evenly distributed across the target, so specific spectral bands L_C can reflect distance information to some extent.

Based on the above analysis, RGB_L and L_C are used as inputs to construct a hybrid deep-learning network from the perspectives of color and distance. The dual-modal feature extraction module (DFEM) and the semantic feature module are incorporated.

1. Dual-Modal Feature Module: DFEM mainly fuses visual information and distance information. The Coordinate Attention (CA) module is introduced to associate local pixel class information with global class information, which promotes the extraction of corresponding object features. The structure of the RGB feature extraction module and the distance information extraction module is consistent. Convolutional layers are used to extract features, while Batch Normalization (BN) and ReLU are applied to improve model stability and prevent overfitting. The block consists of two identical convolutional neural networks, passing through 3×3 deep convolution layers, batch normalization layers, and ReLU activation functions to enhance model stability and prevent overfitting. Max pooling operations are used to adjust the feature size.

   Channel concatenation is used to fuse the features from different modules, constructing a CA module to enhance the model's expressive capability and performance for multimodal input data. Transposed convolution operations are then applied to adjust the size of the output image, allowing the fusion of features from different modalities and hierarchical levels. Finally, two convolution layers are used to further extract features, as shown in Fig. 2. Let $X = [x_1, x_2, \ldots x_c]$ represent the input, and the corresponding output for C channels is:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{5}$$

where H and W represent the height and width of the input, respectively. Kernels of size (1, W) and (H, 1) are constructed to encode each channel along the horizontal and vertical directions. The corresponding outputs for the C-th channel with height h and width w are given as follows:

$$\begin{cases} Z_c^h(i) = \frac{1}{W} \sum_{j=1}^{W} x_c(h,j) \\ Z_c^w(i) = \frac{1}{H} \sum_{i=1}^{H} x_c(i,w) \end{cases} \tag{6}$$

After concatenating Z^h and Z^w, a 1×1 convolution $F_{1\times1}$ and activation function δ are applied to obtain the feature map:

$$f = \delta(F_{1\times1}(Concat(Z^h, Z^w))) \tag{7}$$

Along the spatial dimensions, it is divided into f^h and f^w, and then dimensionality is increased using $F_{1\times1}$. The final attention vector is obtained through σ:

$$\begin{cases} g^h = \delta(F_{1\times1}(f^h)) \\ g^w = \delta(F_{1\times1}(f^w)) \end{cases} \tag{8}$$

The final output y incorporates the information from x, g^h, and g^w:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

The positional attention mechanism incorporates multimodal feature information while capturing position-sensitive information, enabling the model to accurately locate the categories.

2. Semantic Feature Extraction Module: After obtaining information from DFEM, it needs to be abstracted into semantic information. This module consists of a backbone network and a Hierarchical Perceptual Network (HPN).

The backbone network is optimized based on MobileNetV3, considering the lightweight nature of the network. MobileNetV3, constructed through Neural Architecture Search (NAS), is essentially an optimization of MnasNet. (1) The network structure was analyzed, and redundant components were removed. (2) The channel number of bottleneck layers was optimized to fully extract information. (3) The Squeeze-and-Excitation (SE) module was enhanced to improve network efficiency.

To further optimize performance, some bottleneck blocks in MobileNetV3 were replaced with Dual-path Bottleneck Blocks (DBB), which have lower parameter counts. The constructed E-MobileNetV3 features a 16-layer structure, as shown in Fig. 3, consisting of a series of Bneck and DBB layers with gradually increasing channel numbers. The DBB structure, depicted in Fig. 4, has two branches. Different convolution kernels are used to capture features at various scales. Branch aaa consists of two 1×1 convolutions and one (3, 3) depthwise (DW) convolution. Branch bbb consists of two (1, 1) convolutions and one (5, 5) DW convolution. The output

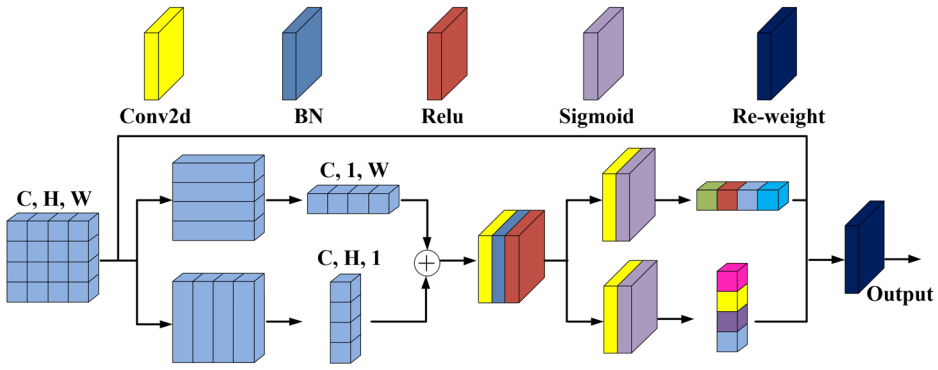


Fig. 2 | Coordinate attention (CA).

features from the two paths are concatenated and then fused using Pointwise Convolution (PWConv).

Since global and local information can enhance the feature extraction of the target, an attention mechanism is constructed based on the output of E-MobileNetV3.

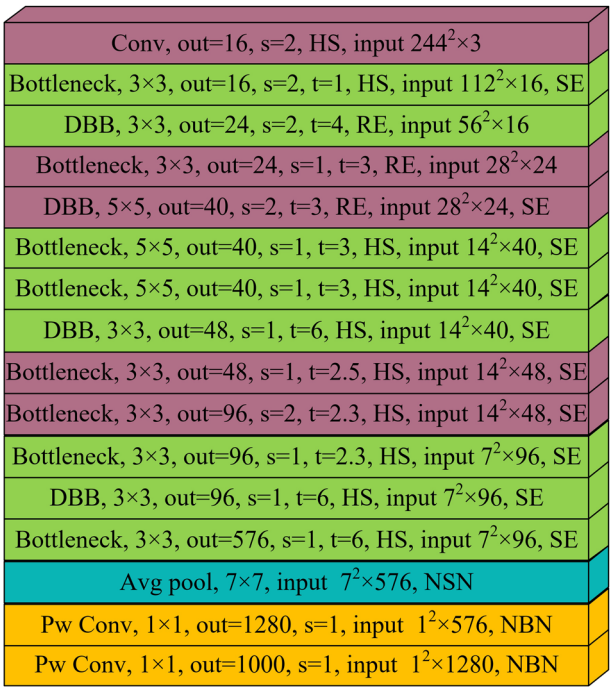


Fig. 3 | MobileNetV3 network structure HS (H-Swish), RE (ReLU), S represents the stride, and NBN indicates no BN operation.

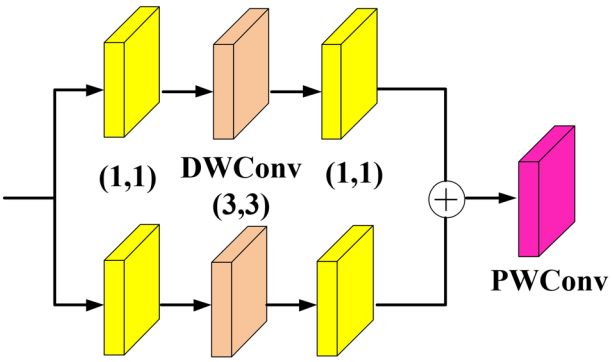


Fig. 4 | Dual-path bottleneck block structure.

The traditional attention mechanism is shown in Fig. 5a. The feature matrices are Q, K, $V \in R^{H\times W\times C1}$, where H, W, and $C1$ represent the height, width and dimension of the features, respectively. The attention mechanism corresponds to three feature convolution matrices W^Q, W^K, and $W^V \in R^{C1\times C}$, where C represents the number of channels. The outputs q, k, $v \in \mathrm{R}^{H\times W\times C}$, with $q = QW^Q$, $k = KW^K$, and $v = VW^V$. The output is:

$$Z_i = \sum_{j=1}^{H\times W} \alpha_{ij} v_j \tag{10}$$

$$e_{ij} = \frac{q_i k_j^T}{\sqrt{C}} \tag{11}$$

This method considers the information of key focus areas but ignores the contextual information of the pixels within the scene. Therefore, this paper refines the attention mechanism by optimizing it with contextual information and target distance information as shown in Fig. 5b. First, additional contextual information is introduced:

$$e_{ij} = \frac{(q_i D)k_j^T}{\sqrt{C}} \tag{12}$$

$$D = diag(\sigma(W_1 W_0(AvgPool(Q))) + W_1(W_0(MaxPool(Q)))) \tag{13}$$

where W_0 and W_1 are parameters, AvgPool and MaxPool represent average pooling and max pooling, respectively, and $diag(\cdot)$ denotes matrix diagonalization. positional information is incorporated into the model:

$$e_{ij} = \frac{(q_i D)k_j^T + q_i r_{ij}^T}{\sqrt{C}} \tag{14}$$

where r_{ij} represents the relative position encoding:

$$r_{ij}(u) = \max(-\xi, \min(u, \xi)) \tag{15}$$

where ξ represents the maximum pixel-level distance. The function $g(x)$ maps the distance to an integer within a finite set, reducing computational costs.

Traditional networks rely solely on upsampling to obtain high-level semantic features, which often leads to a loss of detailed information. To address this, a Multi-Scale Feature Aggregation (MSFA) module is constructed. This module fuses two low-resolution features from the encoder with high-resolution features. The concatenated feature map is processed through one (1, 1) convolution block and two (3, 3) convolution blocks, enhancing the model's feature mapping capability to reconstruct the resolution of the input image. By employing upsampling, the feature map is restored to provide better contextual and detailed information. This approach effectively improves the model's ability to segment small objects within the image.

Fig. 5 | Attention mechanism. a Tradition attention mechanism. **b** Scene-aware attention (SAA).

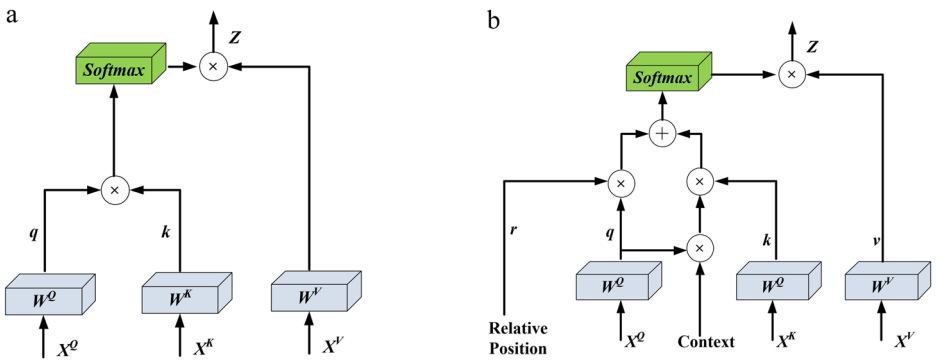


Fig. 6 | Collection equipment.

Results

The field collection equipment for the Terracotta Warriors is shown in Fig. 6. The hyperspectral spectrometer is based on the Dyson mechanism, and its spectrum range is from visible light to near infrared with a line scanning method including 256 spectrum segments. White light source is used for illumination. In total, 51 sets of large-scale hyperspectral data were collected on-site using the hyperspectral spectrometer. Since the algorithm in this paper is based on the deep-learning framework, the data are divided into 1020 sets of small-sized hyperspectral databases according to the unified image resolution. The area where the Terracotta Warriors are located is marked by professionals as the gold standard. The ratio of training data to test data is 2:1. Training data and test data do not overlap to ensure the effectiveness of the algorithm. The experiments were conducted on a server configured with an Intel(R) Xeon(R) Platinum 8260 CPU, eight NVIDIA GeForce RTX 3090 GPUs, and 512 GB of memory, running on the Ubuntu 18.04.6 operating system.

Representative hyperspectral data of the Terracotta Warriors are shown in Fig. 7. It can be observed that the hyperspectral data exhibit a certain level of continuity, and each spectral band demonstrates distinct characteristics. Data 1 includes the torso and partially exposed structures of the Terracotta Warriors. Data 2 features two separated regions of a damaged Terracotta Warrior, with the middle section missing. Data 3 includes the head and other regions of the Terracotta Warriors. Data 4 contains images of the legs of the Terracotta Warriors.

Three spectral bands were extracted from the hyperspectral data to form pseudo-color images, as shown in Fig. 8. Different combinations of spectral bands produce pseudo-color images with varying characteristics, emphasizing different types of information. The pseudo-color image formed by the (0, 8, 16) bands appears darker and lacks sufficient detail representation. The pseudo-color image formed by the (0, 24, 104) bands appears greener, with improved representational ability. The pseudo-color image formed by the (8, 16, 48) bands appears yellowish, showing enhanced ability to represent detailed texture information. The pseudo-color image formed by the (19, 36, 55) bands is the closest to the RGB spectrum, providing the best simulated visual effect. Therefore, this band combination is selected as the input for the algorithm and is denoted as M_{RGB}.

Different spectral bands were selected to simulate distance information, as shown in Fig. 9. Where the z axis represents the relative distance. The 8, 48, and 72 bands could reflect distance information to some extent, but their representation showed large fluctuations. In contrast, the 32nd band exhibited smoother variations and showed significant decreases in the presence of cracks, consistent with visual observations. Therefore, the 32nd band is selected to simulate distance information and is used as the input for the algorithm, denoted as M_L.

Parameter experiments

In order to verify the relationship between the number of RSFEM and their effect, the experiment result as shown in Fig. 10, a single RSFEM has a faster convergence speed and a higher loss value. Due to the increase in the number of parameters, the convergence speed of the 2 RSFEM decreased somewhat, and the loss value decreased significantly. Further increasing the number of parameters of the 3 RSFEM results in a further decrease in the convergence rate and a continued reduction in the loss value. For 4 RSFEM, the number of parameters increases and the convergence rate is slow, but the loss value is close to that of 3 RSFEM. Considering both convergence speed and performance, we selected 3 RSFEM to carry out follow-up experiments.

In this study, the algorithm divides the data into three categories: hyperspectral data, simulated pseudo-color data, and simulated distance data. Corresponding feature extraction models are established: the hyperspectral-based feature extraction model and the visual-and-distance-based feature extraction model. Since the collection site can be considered as comprising the Terracotta Warriors and other regions, the focus is placed on the regions containing the Terracotta Warriors. To validate the performance of each module, we introduce metrics AOM (Area Overlap Measure), AVM (Area oVersegentation Measure), AUM (Area Undersegmentation Measure) and CM (Combination Measure)[29] for evaluation. The corresponding relationships are shown in Fig. 11:

$$\begin{cases} AOM = \frac{A \mid B}{A \cup B} \\ AVM = \frac{A-B}{A} \\ AUM = \frac{B-A}{B} \\ CM = \frac{1}{3}\{AOM + (1 - AVM) + (1 - AUM)\} \end{cases} \tag{16}$$

where A represents the ground truth, and B represents the segmentation results of the algorithm. AOM and CM are directly proportional to the algorithm's performance, while AVM and AUM are inversely proportional to the algorithm's performance.

The effects of different inputs on the algorithm's performance are shown in Table 3. M_{HIS} uses the hyperspectral-based feature extraction model for experiments. M_{RGB} and M_L use the visual-and-distance-based feature extraction model for experiments. When using only M_{RGB} or M_L as input, the single-path input within the visual-and-distance-based feature extraction model is tested. Type 1 uses M_L as input, constructing the model solely from the perspective of distance. The representation capability is insufficient, resulting in poor performance. Type 2 uses M_{RGB} as input, constructing the model from the perspective of vision. While this provides some effectiveness, it still lacks adequate representation. Type 3 uses M_{HIS} as input, extracting features from HIS data, which increases the useful information and leads to improved performance. Type 4 uses M_{RGB} and M_L as inputs, fusing visual and distance information, resulting in further improvement in performance. Type 5 uses M_{HIS} and M_{RGB} as inputs, fusing HIS data features and visual information, leading to enhanced performance. Type 6 uses M_{HIS} and $M_L$as inputs, fusing HIS and distance information, achieving better performance. Type 7 fuses HIS, visual, and distance information, achieving the best performance.

Algorithm comparison results

To verify the effectiveness of the proposed algorithm, we compared it with mainstream algorithms, as shown in Table 4. The proposed algorithm adopts the $M_{HIS\text{-}RGB\text{-}L}$ approach. SVM[30] extract hyperspectral image features, and classify the features through Support vector machines to achieve the segmentation of the Terracotta Warriors from background. However, this method did not consider the difference of near-distance light imaging, resulting in poor segmentation effect, Its CM reached 0.79. Graph-FCN[31] initializes a graph model with a fully convolutional network (FCN) for image semantic segmentation. It transforms image grid data into graph-structured data and formulates the semantic segmentation task as a graph node classification problem, enabling segmentation of Terracotta Warrior images. Its CM reached 0.81. SVM + MRF[32] based on MRF and SVM,

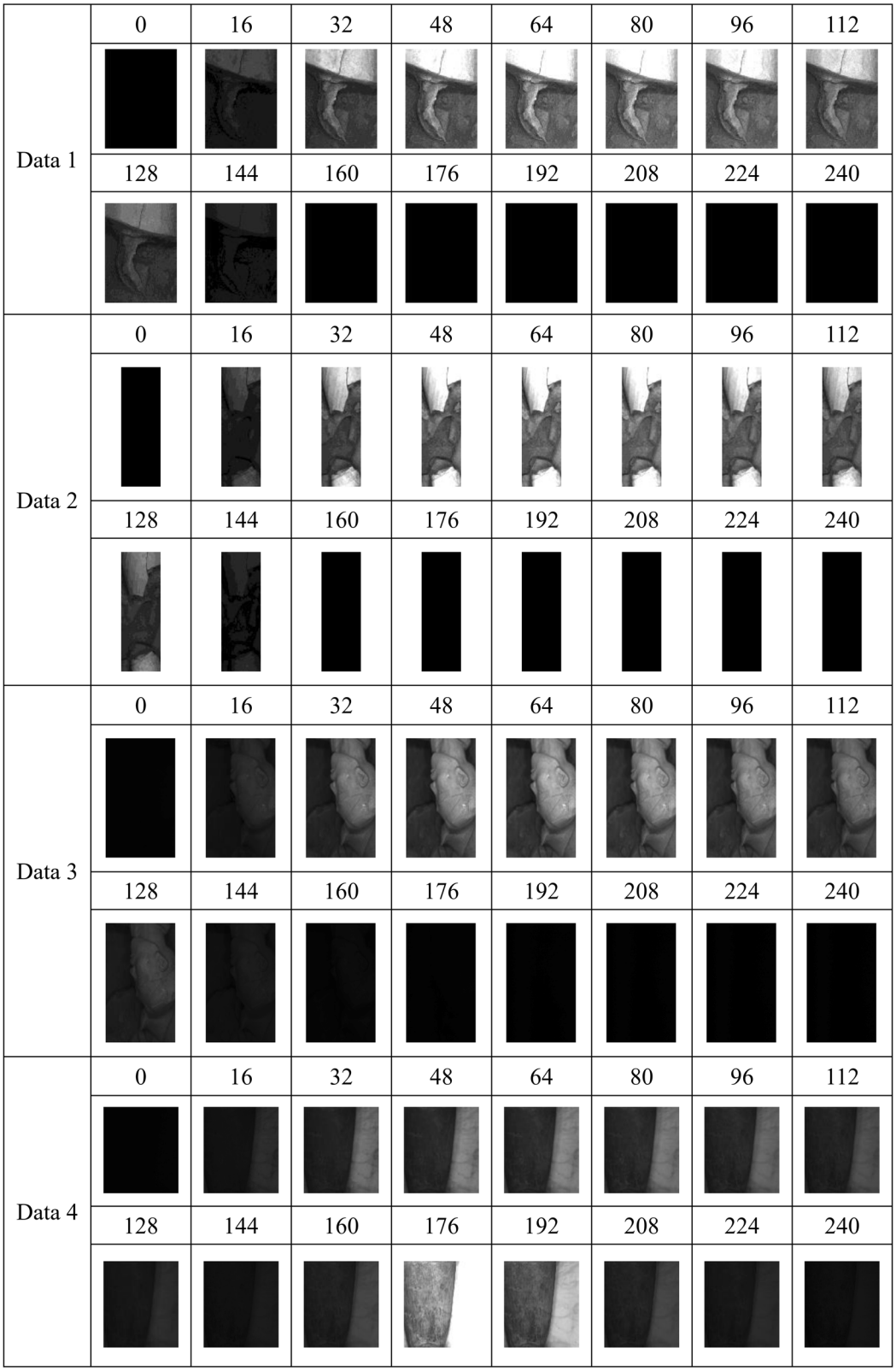
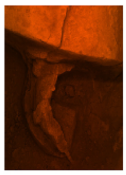
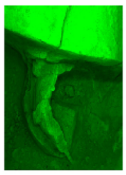
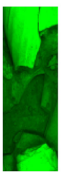
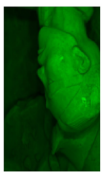
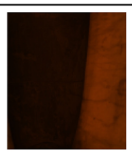
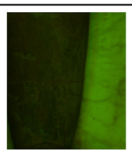
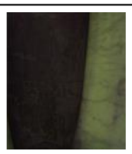
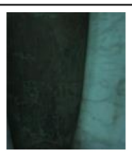


Fig. 7 | Hyperspectral data of the Terracotta Warriors.

which considered the reciprocity among pixels. Its *CM* reached 0.83. PSPNet[33] uses ResNet and Dilated Network structures for feature extraction. By computing the loss and employing backpropagation, it learns and adjusts parameters to perform Terracotta Warrior extraction, Its *CM* reached 0.85. Deep-Unet[34] is an improved version of the traditional Unet. It preprocesses data to highlight specific features, achieving Terracotta Warrior extraction. Its *CM* reached 0.84. Swin-Unet[35] builds on the traditional Unet by incorporating a hierarchical Swin Transformer with shifted windows as the encoder to extract contextual features, achieving Terracotta Warrior extraction. Its *CM* reached 0.87. CRF-CNN[36] leverages prior knowledge and

Fig. 8 | Pseudo-color images synthesized from different spectral bands.

(R,G,B)	(0, 8, 16)	(0,24, 104)	(80, 16, 48)	(19,36,55)
Data 1				
Data 2				
Data 3				
Data 4				

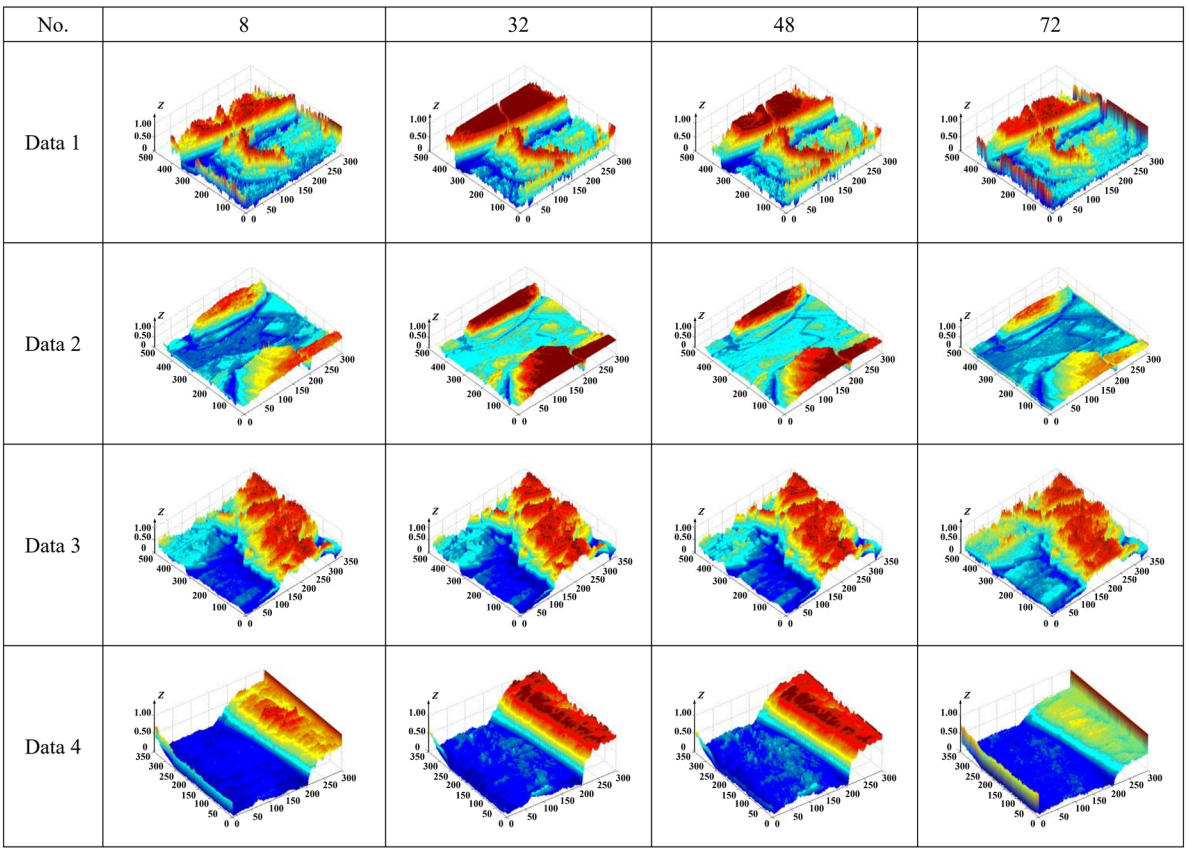


Fig. 9 | Distance information corresponding to different spectral bands.

uses CNN features to extract information. It employs a semi-supervised fully connected Conditional Random Field (CRF) to further refine the features, achieving Terracotta Warrior extraction. Its *CM* reached 0.91. The proposed algorithm constructs a model from the perspectives of hyperspectral, RGB, and distance data, achieving the best performance with a *CM* of 0.93.

The performance of different algorithms on the representative dataset is shown in Fig. 12a represents the results of expert annotations, where yellow indicates the Terracotta Warrior regions, and blue represents other regions. Figure 12b shows the results of expert annotations incorporating distance information, where both yellow and red indicate Terracotta Warrior regions. Red indicates regions closer to the hyperspectral sensor, while yellow indicates regions farther from it. The results of current mainstream algorithms are shown in Fig. 12c–g. Since these algorithms do not incorporate distance information, only blue and yellow regions are represented. The Graph-FCN algorithm results are shown in Fig. 12c. Due to its focus solely on global information while neglecting detail, misdetections occur in the connection areas. The CRF-CNN algorithm results are shown in Fig. 12d. This algorithm builds a model based on prior knowledge, achieving relatively good performance. However, the need for manual interaction during computation affects segmentation efficiency. PSPNet algorithm results are shown in Fig. 12e. The Deep-Unet algorithm results are shown in Fig. 12f. While the preprocessing module enhances useful information, it also suppresses some detail, particularly boundary information, resulting in smaller extracted Terracotta Warrior regions overall. The Swin-Unet algorithm results are shown in Fig. 12g. By using a hierarchical Swin Transformer with shifted windows as the encoder to extract contextual features, the algorithm achieves good performance. However, misdetections occur in small boundary areas, as observed in data 4. The annotation results of the proposed algorithm are shown in Fig. 12h, achieving the highest similarity to expert annotations. Since this algorithm simulates distance information, it can represent certain distance characteristics. As shown in Fig. 12i, the algorithm can separate different Terracotta Warrior regions effectively.

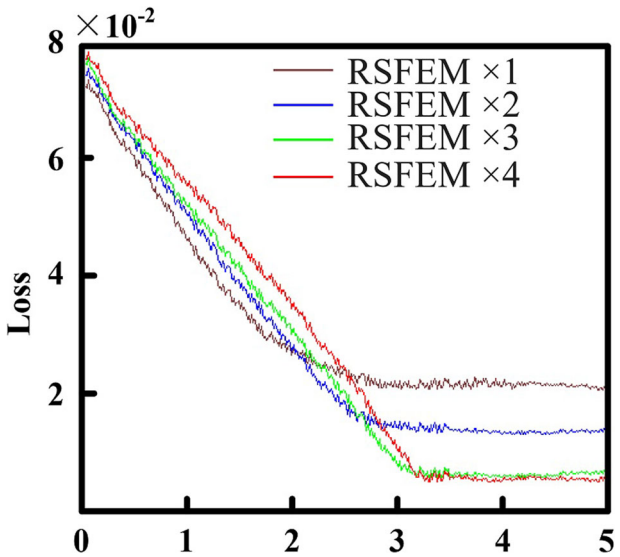


Fig. 10 | Training loss curve.

Discussion

The Terracotta Warriors of the Qin Shi Huang Mausoleum, one of the world's eight wonders, hold significant historical, archeological, and research value. Through in-depth research, this study proposed a multi-dimensional extraction algorithm for the Terracotta Warriors based on hyperspectral imaging. (1) A hyperspectral-based feature extraction model was constructed to reduce data volume while deeply exploring spectral information. (2) A visual-and-distance-based feature extraction model was developed to simulate visual and distance information, enhancing perceptual data extraction. (3) A parallel deep network was built to integrate spectral, simulated visual, and distance information, enabling accurate extraction of the Terracotta Warriors.

However, the study still faces certain limitations that require further research. (1) Due to the high historical and cultural value of the Terracotta Warriors, the availability of large datasets for research is limited. Future efforts could focus on data simulation to expand the database. (2) Limited by the current use of equipment parameters and focal length, the effectiveness of the algorithm is only proved in the range of 08–1.2 M. Subsequently, different spectrograph algorithms can be studied.

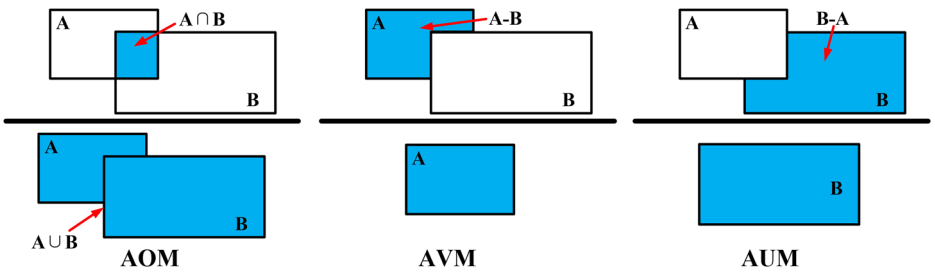
Table 3 | Performance comparison experiments

Type	M_{HIS}	M_{RGB}	M_L	AOM	AVM	AUM	CM
1		√		0.75	0.21	0.24	0.77
2	√			0.78	0.18	0.21	0.80
3	√			0.82	0.15	0.19	0.83
4	√	√		0.85	0.13	0.16	0.85
5	√	√		0.87	0.14	0.14	0.86
6	√	√		0.90	0.12	0.13	0.88
7	√	√	√	0.96	0.08	0.10	0.93

Table 4 | Performance of different algorithms

Algorithm	AOM	AVM	AUM	CM
SVM[30]	0.79	0.23	0.19	0.79
Graph-FCN[31]	0.80	0.20	0.17	0.81
SVM + MRF[32]	0.82	0.16	0.17	0.83
PSPNet[33]	0.85	0.15	0.14	0.85
Deep-Unet[34]	0.82	0.18	0.13	0.84
Swin-Unet[35]	0.87	0.13	0.12	0.87
CRF-CNN[36]	0.93	0.10	0.11	0.91
Ours	0.96	0.08	0.10	0.93

Fig. 11 | Correspondence of evaluation metrics.



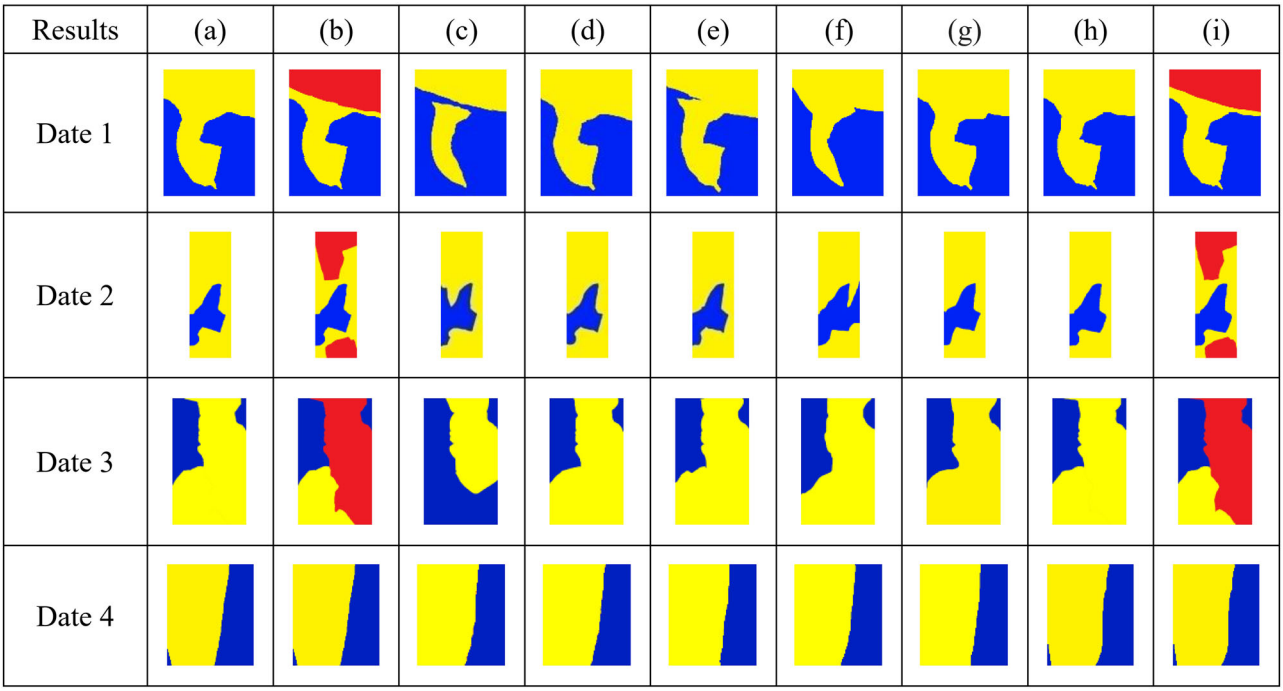


Fig. 12 | Terracotta Warrior extraction results. a Classification annotation result. **b** Annotated results with distance information. **c** G-FCN, **d** CRF-CNN, **e** PSPNet, **f** Deep-Unet, **g** Swin-Unet, **h** Ours, **i** Ours-With distance.

Data availability

The data used to support the findings of this study are available from the corresponding author upon request.

Received: 14 January 2025; Accepted: 29 May 2025;
Published online: 13 June 2025

References

1. Liu, B. et al. *Bacillus bingmayongensis* sp. nov., isolated from the pit soil of Emperor Qin's Terra-cotta warriors in China. *Antonie Van. Leeuwenhoek* **105**, 501–510 (2014).
2. Cheng, Y. & Liu, X. A method of extracting outer eye corners of terra cotta warriors based on point cloud data. In *Third International Conference on Image Processing and Intelligent Control (IPIC 2023)* 9–15 (SPIE, 2023).
3. Hu, Y., Wang, J. & Lan, D. Statistical analysis of the differences of head and face features between terracotta warriors and modern multi ethnic groups based on 3D information extraction. *IOP Conf. Ser.: Earth Environ. Sci.* **783**, 012096 (2021).
4. Liu, J. et al. AMS-Net: an attention-based multi-scale network for classification of 3D terracotta warrior fragments. *Remote Sens.* **13**, 3713 (2021).
5. Hu, Y., Geng, G., Li, K., Guo, B. & Zhou, P. Self-supervised segmentation for Terracotta Warrior point cloud (EGG-Net). *IEEE Access* **10**, 12374–12384 (2022).
6. Lei, Z. et al. A complete methodology for structural finite element analysis of Terracotta Sculptures based on a 3D scanned model: a case study on the Terracotta Warriors (China). *Stud. Conserv.* **68**, 761–772 (2023).
7. Gao, Z., Zheng, H. & Yu, Y. Holographic three-dimensional imaging of Terra-Cotta Warrior model using fractional Fourier transform. *J. Imaging* **5**, 67 (2019).
8. Sheng, Y. Facial recognition and classification of Terracotta Warriors in the mausoleum of the first emperor using deep learning. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **10**, 205–212 (2024).
9. Yang, K., Cao, X., Geng, G., Li, K. & Zhou, M. Classification of 3D terracotta warriors fragments based on geospatial and texture information. *J. Vis.* **24**, 251–259 (2021).
10. Li, F., Gao, Y., Candeias, A. & Wu, Y. Virtual restoration system for 3D digital cultural relics based on a fuzzy logic algorithm. *Systems* **11**, 374 (2023).
11. Rui, X., Zhang, X. & Wang, R. Application of the multi-feature splicing technology based on residual network identification. In *2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA)* 381–384 (IEEE, 2021).
12. Wang, C. et al. Microbial diversity in earthen site of exhibition Hall of pit no. 1 at the Terracotta Warriors Museum in Emperor Qinshihuang's Mausoleum Site Museum and its correlation with environmental factors. *Front. Microbiol.* **15**, 1378180 (2024).
13. Wang, J. et al. Microscopic imaging technology assisted dynamic monitoring and restoration of micron-level cracks in the painted layer of Terracotta Warriors and horses of the Western Han Dynasty. *Polymers* **14**, 760 (2022).
14. Cui, Y. et al. Hidden-information extraction from layered structures through terahertz imaging down to ultralow SNR. *Sci. Adv.* **9**, eadg8435 (2023).
15. Gou, Y. et al. Enhancement of aluminum phosphate adhesion performance by nano-clay for terracotta figurine restoration. *Int. J. Adhes. Adhesives* **132**, 103685 (2024).
16. Xia, Y., Chang, B., Yu, C., Gu, Z. & Luo, X. Evaluating the impact threshold of soluble salts on preservation of earthen site in Emperor Qin Shi Huang's mausoleum. *Indoor Built Environ.* **34**, 388–401 (2024).
17. Zhang, X., Yuan, S., Wu, Y., Guo, B. & Han, J. Research on the corrosion of bronze weapons from the Pits of the Terracotta Warriors. *MRS Online Proc. Libr.***1319**, mrsf10–mrsf1319 (2011).
18. Qiu, S., Ye, H. & Liao, X. Coastal zone extraction algorithm based on multilayer depth features for hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **61**, 5527315 (2023).
19. Zeng, Z. et al. Virtual restoration of ancient tomb murals based on hyperspectral imaging. *Herit. Sci.* **12**, 1–18 (2024).

20. Han, D., Ma, L., Ma, S. & Zhang, J. Discovery and extraction of surface painted patterns on the cultural relics based on hyperspectral imaging. *J. Phys. Conf. Ser.* **1237**, 032028 (2019).
21. Feng, T. & Wang, Y. Evaluation method of authenticity of ceramic and porcelain cultural relics based on spectral analysis. In *AOPC 2022: Optical Sensing, Imaging, and Display Technology*, Vol. 12557, 335–341 (SPIE, 2022).
22. Wei, C., Li, J. & Liu, S. Applications of visible spectral imaging technology for pigment identification of colored relics. *Herit. Sci.* **12**, 321 (2024).
23. Qiu, S., Zhang, P., Li, S. & Hu, B. Extraction and analysis algorithms for Sanxingdui cultural relics based on hyperspectral imaging. *Comput. Electr. Eng.* **111**, 108982 (2023).
24. Peng, J. et al. Mining painted cultural relic patterns based on principal component images selection and image fusion of hyperspectral images. *J. Cult. Herit.* **36**, 32–39 (2019).
25. Liu, S., Wei, C., Li, M., Cui, X. & Li, J. Adaptive superpixel segmentation and pigment identification of colored relics based on visible spectral images. *Herit. Sci.* **12**, 350 (2024).
26. Wang, S. et al. Virtual restoration of ancient mold-damaged painting based on 3D convolutional neural network for hyperspectral image. *Remote Sens.* **16**, 2882 (2024).
27. Pan, N. & Hou, M. The extraction and fusion of faint mural based on feature transform of hyperspectral images. In *2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA)* 161–164 (IEEE, 2016).
28. Peng, J. et al. A relic sketch extraction framework based on detail-aware hierarchical deep network. *Signal Process.* **183**, 108008 (2021).
29. Qiu, S., Jin, Y., Feng, S., Zhou, T. & Li, Y. Dwarfism computer-aided diagnosis algorithm based on multimodal pyradiomics. *Inf. Fusion* **80**, 137–145 (2022).
30. Pal, M. & Foody, G. M. Feature selection for classification of hyperspectral data by SVM. *IEEE Trans. Geosci. Remote Sens.* **48**, 2297–2307 (2010).
31. Lu, Y., Chen, Y., Zhao, D. & Chen, J. Graph-FCN for image semantic segmentation. In *International Symposium on Neural Networks* 97–105 (Springer International Publishing, 2019).
32. Tarabalka, Y., Fauvel, M., Chanussot, J. & Benediktsson, J. A. SVM- and MRF-based method for accurate classification of hyperspectral images. *IEEE Geosci. Remote Sens. Lett.* **7**, 736–740 (2010).
33. Zhu, X., Cheng, Z., Wang, S., Chen, X. & Lu, G. Coronary angiography image segmentation based on PSPNet. *Comput. Methods Prog. Biomed.* **200**, 105897 (2021).
34. Singh, N. & Nongmeikapam, K. Semantic segmentation of satellite images using deep-unet. *Arab. J. Sci. Eng.* **48**, 1193–1205 (2023).
35. Cao, H. et al. Swin-unet: unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision* 205–218 (Springer Nature Switzerland, 2022).
36. Maggiolo, L., Marcos, D., Moser, G., Serpico, S. & Tuia, D. A semisupervised CRF model for CNN-based semantic segmentation with sparse ground truth. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2021).

Acknowledgements

This work is supported by Shaanxi key research and development plan (No. 2024SF-YBXM-678). Shaanxi Province key industrial innovation chain (No. 2023-ZDLGY-45). Light of West China (No. XAB2022YN10). the Key research and development plan of Ministry of Science and Technology (No. 2023YFF0906200).

Author contributions

Shi Qiu was in charge of conducting algorithmic research and performing experimental analysis. Pengchang Zhang modified the algorithm's settings. Siyuan Li gathered the necessary data. Bingliang Hu prepared and refitted the equipment for acquiring hyperspectral data. All authors were involved in the conceptualization and design of the research, as well as in the analysis and interpretation of the data. They also contributed to the writing of the work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Pengchang Zhang.