Available Online: 18 November 2026 DOI: 10.54254/2753-7080/2026.39245

Interpretive model of emotional expression in sculptural works based on visual-affective semantic analysis

An Zhao

School of Humanities and Social Science, Xi'an Jiaotong University, Xi'an, China

zhaoan02529@outlook.com

Abstract. The expression of emotions in sculptures is conveyed through structural forms, spatial relationships, material characteristics, and symbolic elements. However, traditional artistic analysis mainly relies on expert interpretation, making emotional expression difficult to quantify. With the development of computer vision and multimodal Artificial Intelligence (AI), deep learning approaches provide new opportunities for visual art understanding. Nevertheless, existing methods still face challenges in establishing connections between visual features and abstract emotional semantics, as well as providing interpretable prediction processes. To address these issues, an interpretive model of emotional expression in sculptures based on visual-affective semantic analysis is proposed. A public sculpture image dataset containing 1,800 instances is constructed. A 50-layer Residual Network (ResNet50) is employed to extract morphological and textural features, while Contrastive Language–Image Pre-training (CLIP) is used to establish associations between visual representations and affective semantics. An attention-based fusion mechanism integrates visual and semantic features for emotion classification. Gradient-weighted Class Activation Mapping (Grad-CAM) and semantic ablation methods are further applied to analyze the decision-making process of the model. Experimental results show that the visual-semantic fusion model achieves an accuracy of 82.59% and a Macro-averaged F1 score (Macro-F1) of 80.73% on the test set, outperforming the ResNet50 visual model with 76.30% accuracy and 75.67% Macro-F1, as well as the CLIP semantic model with 73.85% accuracy and 72.58% Macro-F1. Interpretation analysis demonstrates that the model mainly focuses on essential artistic elements, including posture, contour, and spatial structure, with strong consistency with expert evaluations. This research provides an effective artificial intelligence framework for sculptural emotion analysis and offers potential applications in digital museums, art education, intelligent retrieval systems, and computational cultural heritage studies.

Keywords: sculptural emotional expression, visual emotion analysis, semantic features, multimodal fusion, explainable artificial intelligence

1. Introduction

Sculptures combine structural form, space, material, and cultural symbolic meanings in using form, volume, texture, and space arrangement to express emotions. Nevertheless, emotional understanding is always subjective because of the different personal experiences, cultural backgrounds, and aesthetic perception [1].

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Traditionally, artistic analysis depends highly on human experts' interpretation, so that quantitative analysis is difficult. Thanks to the recent developments in computer vision and deep learning, there is now the possibility of visual analysis and emotion recognition by using datasets which can uncover the relationship between the visual cues and the emotion perception [2]. Nevertheless, previous works mainly concentrated on natural images with little research done on the sculptures, and lacking explanation for the effects of visual cues on emotion perception. Thus, a visual-affective semantic framework is proposed to incorporate visual cues and emotional concepts.

2. Literature review

2.1. Computational representation of sculptural form

The emergence of computer vision allows artworks to be converted into computational representations by means of feature learning technologies. The early works of image analysis were focused mostly on the extraction of low-level features like colors, edges, and textures; however, the emotional meanings contained in artworks usually exist in the global structure, local form, and visual relation. The previous studies on the visual emotion analysis proved that composition, relations of objects, and visual components affect the emotional perception and create links between image features and affective classes [3]. In the case of sculpture, Two-Dimensional (2D) images do not capture completely the Three-Dimensional (3D) properties of sculptures, which makes the orientation of posture, variations of contour, proportion of space, and materials used in sculptures significant elements for the emotional perception of them. Additional researches showed that the analysis of artistic images should take into account the inherent formal characteristics of artworks to avoid dependency of the models on such irrelevant features as background, lighting, and conditions of photographing [4]. It leads us to conclude that the computational representations of sculpture should combine deep visual features and art-specific structural descriptions.

2.2. Visual-affective semantic mapping

Visual-affective semantic mapping emphasizes how visual input becomes emotional meaning in order to become a key link between computational approaches and artistic understanding. Traditional visual emotion analysis methods depend on colors, composition, and object classes, while sculptures convey their emotions through postures, movement direction, spatial tensions, and symbolic elements. Thus, specific visual attributes may not be enough to interpret abstract emotional meanings of sculptural pieces. Modern developments of vision-language models make it possible to represent both images and textual concepts in the same semantic space so that the computational approach could understand more detailed emotions and artistic concepts. Some previous studies showed that pretraining with vision-language models helped improve semantic understanding of complex visual content based on image-text associations and increased cross-modal analysis performance [5]. In addition, studies about affective descriptions of visual art proved that emotional language could enhance visual input and perceive affective meanings hidden in the artworks [6]. Hence, the approach of visual-semantic fusion is promising for recognition of emotions in sculptures, while annotation methods and artistic evaluation criteria should be employed to minimize the effect of cultural factors and personal understanding of artworks.

2.3. Explainability in art emotion recognition

It is necessary not only to achieve high accuracy but also to provide the interpretability of the decision process of computational models for emotion recognition in art. While deep learning approaches can learn highly complicated visual patterns, the processes of making decisions usually remain black box, which makes it hard to say if the model predicts emotions based on meaningful artistic features. Explainable Artificial Intelligence (AI) allows improving the interpretability of the models through investigating feature importance and attention maps. The Gradient-weighted Class Activation Mapping (Grad-CAM) method enables generating visual heatmaps that show the areas of an image responsible for the predictions of the model and helps to investigate the relations between visual features and emotional categories [7]. Moreover, further research has revealed that using multimodal attention mechanisms allows the integration of the information at various representation levels and increases the identification of important visual features [8]. Considering the fact that artistic emotion recognition relies on the visual, linguistic, and cognitive information, more advanced explanation approaches are needed. It has been proven that multimodal emotion recognition allows improving the interpretation of complex emotions through the integration of diverse sources of information [9].

3. Experimental methods

3.1. Dataset construction and emotion annotation

The experimental dataset was gathered from openly available art sources, including WikiArt Open Art Database, Europeana digital collection, and The Metropolitan Museum of Art Open Access Collection. The data gathering interval was fixed as 2010 to 2025. Images of sculptures that had detailed information on artists, periods of creation, and artwork categories were chosen to ensure the reliability of the data. In total, 2,460 images of sculptures were collected. After removing duplicates, images of low quality, images of significantly occluded sculptures, and images with significant background interference, 1,800 images were kept. The dataset consisted of stone sculptures, wood sculptures, metal sculptures, and multi-material sculptures. Emotional categories include joy, sadness, anger, fear, calmness, and awe. Emotion labels were annotated by three people having an artistic background according to the unified guidelines of annotation. Additional information about artworks, such as posture openness, body contraction, contour variability, and material features, was also annotated. The majority vote decided the final label for each image. The dataset was divided according to the identity of artworks, consisting of 1,260 training samples (70%), 270 validation samples (15%), and 270 testing samples (15%). It ensures that different viewpoints of one sculpture do not appear in different subsets.

3.2. Visual and semantic feature extraction

After dataset construction, visual and semantic features were extracted through two parallel pathways. The visual pathway employed an ImageNet-pretrained 50-layer Residual Network (ResNet50) model. All images were resized to 224 × 224 pixels, and 2,048-dimensional deep visual features were extracted to represent sculpture contours, local structures, and texture variations. In addition, foreground segmentation was performed to obtain sculpture regions, from which contour curvature, posture direction, and local texture descriptors were calculated as complementary features [10]. To reduce differences among feature scales, normalization was applied to the extracted visual representations, as shown in Equation (1):

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

where z_i represents the normalized feature value, x_i represents the original feature value, μ represents the feature mean, and σ represents the standard deviation.

The semantic pathway used Contrastive Language–Image Pre-training (CLIP) to encode the relationship between sculpture images and affective descriptions. Thirty artistic emotion-related phrases, such as "contracted posture", "upward extension", and "isolated figure", were transformed into text embeddings. The similarity between image embeddings and text embeddings was calculated to construct semantic representations. This process connected visual information with interpretable emotional concepts and provided unified feature representations for multimodal fusion.

3.3. Classification and interpretive evaluation

After obtaining visual and semantic representations, an attention-based fusion mechanism was applied to generate integrated features, which were then classified using a support vector machine. The fusion module dynamically adjusted the contribution of different modalities, allowing the model to focus on more discriminative information according to sculpture categories. The fusion process as shown in Equation (2):

$$F = \alpha V + \beta S \tag{2}$$

where F represents the fused feature representation, V represents the visual feature vector, S represents the CLIP-based semantic feature vector, and α and β represent the learned weights of visual and semantic information.

The fused features were input into a radial basis function support vector machine for six-category emotion classification. Five-fold cross-validation was applied to determine the penalty parameter C and kernel parameter γ. Model performance was evaluated using Accuracy, Macro-averaged F1 score (Macro-F1), Precision, Recall, and Area under the Curve (AUC). To examine whether the model relied on meaningful artistic elements, Grad-CAM was applied to generate visual attention maps [11]. Semantic ablation was further conducted to measure the contribution of individual emotional concepts, as shown in Equation (3):

$$\Delta P = P_{all} - P_{remove} \tag{3}$$

where ΔP represents the semantic contribution score, P_{all} represents the prediction probability of the complete model, and P_{remove} represents the prediction probability after removing a specific semantic feature. The combination of classification performance and interpretability evaluation provided a comprehensive assessment of both recognition capability and decision transparency.

To further evaluate the interpretability of model outputs, two experts with backgrounds in art history and sculpture studies were invited to assess the explanation results of 120 randomly selected testing samples. Each expert independently scored the consistency between model-generated explanations and artistic characteristics on a five-point scale, considering three aspects: the relevance of highlighted visual regions, the correspondence between semantic concepts and sculptural forms, and the overall interpretability of emotional prediction. The final explanation score was calculated by averaging the ratings from the two experts across all evaluated samples.

4. Results

4.1. Classification performance

The classification capability of the presented model was tested using 270 testing images. Three baseline models were used for comparisons, including ResNet50 visual model, CLIP semantic model, and handcrafted

feature model. From Table 1, it can be seen that the visual-semantic fusion model performed best in all approaches. Specifically, the proposed model received 82.59% Accuracy and 80.73% Macro-F1, which outperformed the ResNet50 visual model with 76.30% Accuracy and the CLIP semantic model with 73.85% Accuracy. It means that structural visual information and emotional semantic representation can effectively boost the accuracy of sculpture emotion recognition. Detailed experiments revealed that calmness and awe reached the best F1 scores of 85.12% and 84.36%, since stable postures and upward structures generated distinct visual cues. Fear and anger had higher misclassification rates, as high tension body structures and complicated contours made them easily confuse with each other. Adding CLIP semantic features reduced the misclassification rate of the two classes from 19.6% to 12.8%.

Table 1. Performance comparison of different models in sculpture emotion classification

Model	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	AUC
Handcrafted Features	69.44	68.21	67.85	67.92	0.742
CLIP Semantic Model	73.85	72.90	72.64	72.58	0.781
ResNet50 Visual Model	76.30	75.84	75.52	75.67	0.812
Visual-Semantic Fusion Model	82.59	81.36	80.94	80.73	0.871

4.2. Interpretation consistency

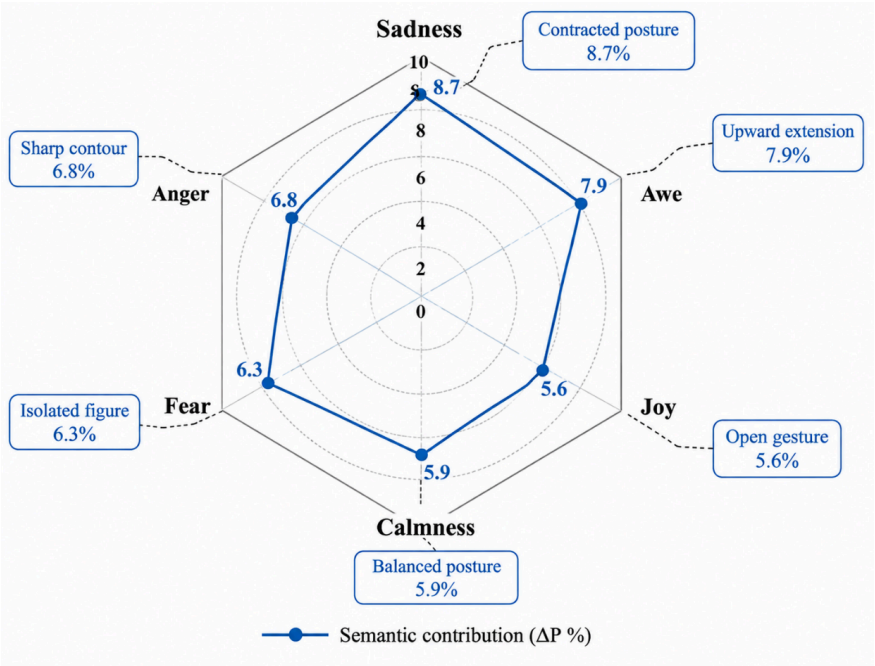


Figure 1. Semantic contribution analysis of emotional concepts in sculpture emotion recognition

In order to examine if the model explanation is consistent with emotional characteristics of sculptures, semantic contribution analysis was performed using 120 test samples. The change of probability caused by the removal of each emotional concept was obtained by Equation (3). As can be seen from Figure 1, different emotional categories were associated with varied contribution of semantic concepts. The semantic concept of "contracted posture" had the highest contribution to sad emotions, causing a mean decrease of 8.7% of probability. "Upward extension" made the greatest contribution to awe emotions, causing a mean decrease of

7.9% of probability after removal. "Sharp contour" made a mean contribution of 6.8% for anger recognition, and "isolated figure" made a mean contribution of 6.3% to fear recognition. The mean expert evaluation score of the explanation results was 4.18 out of 5, suggesting the consistency between the extracted semantic factors and artistic expertise.

5. Discussion

From the experimental results, it can see that by incorporating visual features and affective semantics, emotion recognition accuracy is successfully improved in the sculptural works, which shows that emotion expression in the sculpture is in need of a holistic perception of formal and semantic understanding. The better recognition performance in terms of calmness and awe is mainly because stable posture, vertical extension and openness of space can serve as powerful emotional hints. And the confusion of fear and anger shows the inadequacy of two-dimensional images in expressing complex spatial tension. In this way, the paper enriches the scope of application of computer vision in the field of artistic emotion analysis by expanding the traditional method of visual classification into three-dimensional sculptural expression and its cultural semantics. On the other hand, by jointly modeling visual features and emotional semantics, the connection between formal features and affective concepts can be established, and therefore, the model can analyze the process of emotional expression formation through the analysis of representation of the artwork. Additionally, the combination of Grad-CAM and semantic ablation makes the decision process of the model transparent, and thus, the AI system can identify the influential artistic features in terms of emotion judgment. The alignment of model explanations and experts' evaluations proves that the framework can well capture important visual features such as posture, contours and spatial structures, and provides a new technical approach for digital art analysis and intelligent museum.

6. Conclusion

Emotionality in sculptural art is expressed through visual forms, spaces, and semiotics. In order to deal with the shortcomings of quantitative analysis and model explanations, an interpretative model for visual affective semantic analysis was designed. The visual-affective semantic framework utilizes public sculpture images, visual feature modeling, visual-language semantics encoding, and multimodal fusion in order to recognize emotions, while model explanation techniques help understand the decision-making process in the model. It has been proven experimentally that the visual-semantic framework enhances emotion recognition.

References

[1] Zhang, J., Guo, Y., Zhang, S., & Wang, X. (2024). Training a small emotional vision language model for visual art comprehension. In *European Conference on Computer Vision* (pp. 1–17). Springer Nature Switzerland.

[2] Borriero, A., D'Amelio, A., & Maguolo, G. (2024). Explainable emotion decoding for human and computer vision. In *World Conference on Explainable Artificial Intelligence*. Springer Nature Switzerland.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.

[4] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine*

Learning (pp. 19730–19742). PMLR.

[5] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*(9), 2337–2348.

[6] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barras, F., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, F., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). *Flamingo: A visual language model for few-shot learning*. arXiv. https://arxiv.org/abs/2204.14198

[7] Etesam, Y., Bagheri, S., & Mahoor, M. H. (2024). Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

[8] Zhang, Q., Song, X., Wang, Y., & Liu, X. (2024). Visual prompting in LLMs for enhancing emotion recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

[9] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9650–9660).

[10] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

[11] Ding, T., Yang, J., & Huang, H. (2026). AGER: Attribute-guided emotion reasoning for visual emotion analysis. *IEEE Transactions on Affective Computing*.