Theoretical breakthroughs in the unified representation of aesthetics across modalities from the perspective of computational aesthetics

BU Wei

School of Architecture and Design, Harbin Institute of Technology, Heilongjiang Harbin.

Abstract: Cross-modal aesthetic understanding is a key bottleneck for artificial intelligence to move towards general intelligence. The core challenge lies in how to establish a unified aesthetic representation theory that can connect different sensory modalities. This research is based on the core proposition of computational aesthetics - that is, the mathematical structure of "beauty" can be calculated and represented, and is committed to deepening this proposition in a cross-modal context to solve the key problem of how aesthetic universality manifests in heterogeneous information. This paper systematically analyzes the limitations of existing research in aesthetic semantic alignment, capture of higher-order concepts, and cultural adaptation, and proposes a unified representation theoretical framework centered on the "aesthetic essential space". This framework achieves the stripping and alignment of high-level aesthetic semantics by introducing "aesthetic concept prototypes" as spatial primitives and constructing a "semantic-emotion" collaborative mapping mechanism. Experiments show that the model based on this theory significantly outperforms the baseline in cross-modal aesthetic evaluation, retrieval and generation tasks, verifying its breakthrough value in bridging the "semantic gap" and laying a theoretical foundation for constructing interpretable and generalizable cross-modal aesthetic computing models.

[BU Wei. **Theoretical breakthroughs in the unified representation of aesthetics across modalities from the perspective of computational aesthetics.** *Life Sci J* 2025;22(7):62-67]. ISSN 1097-8135 (print); ISSN 2372-613X (online)**.** http://www.lifesciencesite.com**.** 01. doi:10.7537/marslsj220725.01

Keywords: Computational aesthetics; Cross-modal characterization; Aesthetic essential space; Concept prototype; Semantic alignment

1. Introduction

The core of human aesthetic cognition lies in its common ability that transcends sensory forms. We can intuitively capture the aesthetic unity contained in "poetry is a sound painting, and painting is a silent poem", which closely connects different art forms at the emotional and artistic conception levels. However, this inherent cross-modal aesthetic understanding of human beings constitutes a long-term challenge that artificial intelligence has faced in its process of moving towards higher levels of cognition. Although general multimodal models represented by CLIP[1] have made significant progress in the semantic alignment of vision and language, their representational capabilities are mostly confined to the "descriptive semantics" of the objective world (such as recognizing objects and scenes), while their capture of the subjective and abstract "evaluative semantics" of "what is beauty" appears weak. This limitation makes it difficult for existing models to explain why a tragic symphony and a desolate oil painting can evoke similar aesthetic experiences, and it is even more impossible to achieve high-quality and highly consistent cross-modal aesthetic generation and retrieval.

At its core, the current research predicament stems from the lack of in-depth theoretical knowledge in the field of cross-modal aesthetics. A great deal of work has either been confined to feature optimization within a single modality or relied on shallow associations driven by big data, but has never responded to a core theoretical question: Does there exist an "aesthetic essential space" that transcends specific sensory channels, capable of mapping and measuring the information of heterogeneous modalities based on their inherently unified aesthetic attributes rather than surface physical features? The theoretical vacuum has led to the long-term development of this field being in a state where technology takes the lead, and there are serious deficiencies in the interpretability, generalization ability and cultural adaptability of the models.

To fundamentally break through this bottleneck, this paper systematically proposes a theoretical framework of "unified representation of cross-modal aesthetics" from the perspective of computational aesthetics. The core assumption of this theory is that there exists a structured aesthetic essential space, whose foundation is composed of a series of aesthetic concept prototypes shared across cultures and modalities (such as "sublime", "graceful", and "tragic"). The input of any mode can obtain its coordinates based on aesthetic essence in this space through a specific mapping mechanism.

The fundamental breakthrough of this theory lies in its realization of a paradigm shift from "feature correlation" to "essential abstraction". Specifically, it has for the first time liberated aesthetic representations from the low-level reliance on underlying signals such as color, texture, and musical notes, elevating them to the direct operation and calculation level of abstract concepts like "tragic" and "elegant". To achieve this goal, we have established a "semantic-emotion" collaborative mapping mechanism, which not only captures the content information of aesthetic objects but also deeply integrates the emotional dimensions they evoke, thereby realizing the modeling of a complete aesthetic experience. Furthermore, this framework takes cultural context as an internal regulating variable, enabling the unified representation to dynamically adapt to different aesthetic preferences, thereby possessing cultural robustness that traditional models lack. The work of this article aims to lay an explainable and computable theoretical foundation for building an artificial intelligence that truly "understands beauty".

2. The current situation and theoretical bottlenecks of cross-modal aesthetics research

The development history of cross-modal aesthetics research is closely related to the evolution of representation learning techniques. Looking at its development path, it can be roughly divided into three progressive stages: shallow correlation analysis, deep semantic alignment, and high-level conceptualization attempts. Despite the continuous iteration of technology, each stage has exposed its inherent limitations in approaching the essence of aesthetics, all pointing to the urgent need for a unified theoretical framework.

2.1 Shallow Association Stage: Early Exploration Based on Feature Engineering and Its Limitations

The early research paradigm mainly relied on manually designed features and their statistical analysis. Researchers attempt to establish cross-modal aesthetic connections by calculating the statistical correlations among low-level physical indicators such as image color histograms, texture spectra, and audio spectral features [4]. Although such methods are intuitive, their fundamental flaw lies in simplifying complex aesthetic experiences into a series of measurable physical parameters, thus falling into the predicament of a "feature gap". It is unable to capture the semantic context and subjective experience dimensions on which aesthetic connotations rely, and thus it is difficult to provide a deep explanation for "why certain combinations of colors and sounds are regarded as harmonious", and its application scope and effect are very limited.

2.2 Semantic Alignment Stage: The aesthetic perception of the general model loses focus

With the development of deep learning, especially the emergence of large-scale vision-language pre-trained models represented by CLIP [1], research has entered the stage of deep semantic alignment. This type of model acquires powerful cross-modal semantic matching capabilities through comparative learning on massive amounts of Internet data. However, its success is largely limited to "Descriptive Semantics", that is, the recognition of "What it is", such as accurately associating pictures of "cats" with text. For the "Evaluative Semantics", which is crucial to aesthetic research, namely the judgment of "How it is" or "whether it is beautiful or not", such models show obvious "loss of focus". For instance, the model has difficulty distinguishing the subtle differences in aesthetic attributes between "an elegant cat" and "a comical cat".

The root cause lies in the fact that the aesthetic signals in network data are implicit and noisy, and the model is not explicitly guided to learn the feature dimensions related to aesthetic evaluation, resulting in its representation being insensitive to aesthetic semantics. This is currently the most significant technical bottleneck hindering the understanding of cross-modal aesthetics.

2.3 Advanced Abstraction Stage: Conceptualization attempts and the absence of theoretical frameworks

To overcome the aforementioned limitations, recent studies have begun to attempt to introduce higher-level aesthetic concepts, such as injecting abstract labels like "harmony", "balance", and "sublimity" into knowledge graphs. This marks an important step forward in research towards the essence of aesthetics. However, most current work still remains at the level of "concept labeling", failing to theoretically solve how these discrete and symbolized concepts can be effectively connected with continuous deep representations (i.e., the "symbolic grounding problem"), nor has it clarified how these concepts can serve as an intermediary bridge to unify representations of different modalities.

More crucially, there is a lack of a unified theoretical framework within the field to formally define the intrinsic relationships among these aesthetic concepts and how they are consistently mapped and measured in different modalities. This theoretical vacuum makes it difficult for various studies to communicate with each other, resulting in scattered achievements and preventing the formation of cumulative disciplinary progress.

2.4 Systematic Summary of Core Theoretical Bottlenecks

By sorting out the evolution of cross-modal aesthetics research, three interrelated and progressive core theoretical bottlenecks can be clearly identified, which jointly restrict the breakthrough progress in this field.

The primary and most fundamental bottleneck lies in the fragmentation of representations. Currently, aesthetic information from different modalities such as vision, hearing, and text is mapped into their respective independent and heterogeneous feature Spaces. This leads to the fact that the aesthetic values among different art forms cannot be directly measured and compared on a unified dimension, as if there is a lack of a "universal aesthetic language" that can translate all "sensory dialects". Therefore, constructing a common and structured aesthetic essential space has become a prerequisite for achieving a deep understanding of cross-modal aesthetics.

Secondly, there is the ambiguity of aesthetic semantics. Aesthetic experience is essentially a complex integration of objective content and subjective feelings. For instance, when appreciating a Gothic cathedral, its aesthetic value stems not only from the objective entity of "church" (content semantics), but also from its style semantics of "towering" and "mysterious" (style semantics). Most existing methods confuse the two and fail to clearly separate them, making it difficult for models to accurately understand and generate specific aesthetic intentions, which seriously weakens their interpretability and controllability.

Thirdly, there is the lack of adaptability of the model to dynamic contexts. Aesthetic judgment is not an absolute standard; it is profoundly dependent on dynamic and changing contextual factors such as cultural background, the spirit of The Times, and personal preferences. For instance, the "blank space" artistic conception in Eastern aesthetics and the "complexity" beauty in Western Baroque art represent different cultural aesthetic standards. The existing static models are difficult to effectively capture and integrate these complex and changeable contextual information, resulting in their output results often appearing rigid, lacking cultural sensitivity and personalized expression, which greatly limits their generalization ability and application value in the real world.

3. Construction of a theoretical framework for the unified representation of cross-modal aesthetics

To systematically address the aforementioned theoretical bottlenecks, this chapter proposes a unified representation theory framework centered on the "aesthetic essential space". This framework aims to construct an aesthetic representation system capable of integrating multi-modal information, providing a unified theoretical basis and methodological guidance for cross-modal aesthetic computing.

3.1 Theoretical Foundation: The Aesthetic Essential Space Hypothesis

This framework is based on the core theory of the aesthetic essential space hypothesis. This hypothesis holds that there exists an abstract high-dimensional metric space, whose essential feature lies in its ability to break free from the constraints of specific sensory forms and directly represent the intrinsic unity of aesthetic experiences.

Unlike traditional feature Spaces, space has a unique modal invariance: as long as stimuli of different modalities have similar aesthetic essences, they will form a compact distribution in this space. For instance, an oil painting that showcases the beauty of grandeur, a poem that conveys the artistic conception of grandeur, and a symphony that embodies the momentum of grandeur, although their physical forms are quite different, their positions in space will be highly adjacent. The mathematical structure of space is defined by a set of primitive aesthetic concept prototypes. These prototypes (such as "sublime", "graceful", "tragic", etc.) constitute the basic dimensions of aesthetic experience, equivalent to the coordinate bases of space.

Under this framework, the aesthetic representation of any stimulus can be expressed as a functional combination of these prototypes: This representation method not only realizes cross-modal aesthetic comparison but also provides interpretability of the representation - any complex aesthetic experience can be decomposed into the combination of several basic aesthetic prototypes and their intensity distribution. The innovation of this hypothesis lies in transforming discrete aesthetic concepts into continuous vector representations, thereby providing mathematical tools for quantitative research on aesthetic experiences. By establishing a mapping relationship from specific sensory features to the essence of abstract aesthetics, this framework lays a theoretical foundation for understanding the unity of cross-modal aesthetics.

3.2 Core Mechanism: Semantic-Sentiment dual-path mapping model

To achieve an effective mapping from multimodal input to the aesthetic essential space A, this paper proposes a semantic-emotion dual-path mapping model. The design of this model is based on the fundamental viewpoint of cognitive science, that is, a complete aesthetic experience arises from the synergistic effect of objective cognition and subjective feelings. By establishing two relatively independent yet collaborative processing paths, the model can respectively capture the content information and emotional traits of aesthetic objects, ultimately achieving a unified aesthetic representation.

Semantic pathways, as the cognitive basis of models, are mainly responsible for extracting the objective content semantics of stimuli. This path employs a general model pre-trained on large-scale datasets (such as BERT for text and ViT for images) as the encoder E_sem, which can effectively identify and encode conceptual information such as entities and scenes in the input signal. Take an oil painting depicting a sunset as an example. The semantic pathway will extract key semantic features such as "sunset", "sky", and "oil painting" to form a content semantic vector s. This path ensures that aesthetic representations have a solid semantic foundation and provides content-level consistency guarantees for cross-modal alignment.

The emotional pathway focuses on capturing the subjective emotional experiences triggered by stimuli, including two dimensions: emotional valence (pleasant - unpleasant) and arousal (calm - excited). This path extracts emotion-related features from the input signal through an encoder E_aff specifically trained on the emotion dataset to form the emotion vector e. Emotional pathways can identify and encode emotional traits such as "warmth", "serenity", and "magnificence", infusing subjective experience dimensions into aesthetic representations. The outputs of the two paths are not simply spliced together, but are deeply integrated through a carefully designed collaborative fusion module G. This module maps the semantic vector s and the sentiment vector e to the unified aesthetic essential space A through the learnable parameter Θ_G, forming the final aesthetic representation a = G(s,e;Θ_G). The training objective of the fusion module is to minimize the distance of samples from the same aesthetic concept but different modalities in space A through an optimization algorithm, while maximizing the distance of samples from different aesthetic concepts.

This design ensures that the model can spontaneously discover aesthetic consistency across modalities while maintaining the distinction between different aesthetic concepts. The advantage of the dual-path model lies in maintaining the stability of the semantic foundation while integrating the richness of emotional experience, and it effectively simulates the complexity of human aesthetic cognition. Through the collaborative work of the two paths, the model can go beyond simple content understanding or sentiment analysis and achieve true cross-modal aesthetic representation.

3.3 Implementation Path: Contrastive learning guided by concept prototypes

To address the issue of the lack of detailed aesthetic concept annotation data, this paper proposes a contrastive learning strategy guided by concept prototypes. The core innovation of this method lies in the combination of the discovery process of aesthetic concept prototypes with cross-modal representation learning, enabling the model to spontaneously construct an aesthetic concept system with semantic significance during the learning process.

In terms of initializing the concept prototype, we adopt two complementary schemes. On the one hand, the prototype is randomly initialized as a learnable parameter to enable the model to independently discover meaningful prototype structures from the data. On the other hand, a prior initialization is carried out based on the existing aesthetic knowledge graph, and the conceptual relationships in the existing aesthetic theories are injected into the model as prior knowledge. This dual strategy not only ensures the flexibility of the model but also makes use of existing domain knowledge, providing a good starting point for prototype learning.

The training of the model is jointly guided by two complementary loss functions. Cross-modal alignment loss ensures that samples of different modalities from the same aesthetic description remain adjacent in the aesthetic essential space A. For instance, although "A peaceful piece of music" and "a peaceful painting" have different modalities, due to their shared aesthetic trait of "serenity", their representation vectors should have a high degree of similarity in space A. This loss function prompts the model to capture aesthetic consistency across modalities. Meanwhile, the concept aggregation loss is dedicated to optimizing the internal structure of space A, requiring the representation vectors of samples labeled as the same aesthetic concept to aggregate towards the corresponding concept prototypes, so that each concept prototype naturally becomes the "gravitational center" of this type of aesthetic trait in space. Through this joint optimization mechanism, the model not only achieves alignment of cross-modal representations, but more importantly, spontaneously discovers the conceptual prototypes that serve as the basic units of aesthetic cognition.

This process has the characteristic of theoretical self-verification: if the assumption of the "aesthetic essential space" holds true, then the conceptual prototype should naturally emerge as a stable anchor point for cross-modal alignment. This data-driven verification approach endows the theoretical framework with solid characteristics of being computable and verifiable. This method successfully transforms the abstract theoretical framework into an achievable computational model, maintaining theoretical consistency while ensuring engineering feasibility, providing a complete path from theory to practice for cross-modal aesthetic research. This learning mechanism can not only discover meaningful aesthetic concept structures, but also ensure that the learned representations have good cross-modal generalization ability.

4. Experimental verification and result analysis

To systematically verify the effectiveness and superiority of the cross-modal aesthetic unified representation theory proposed in this paper, we designed and implemented a series of control experiments. The experiment mainly focuses on the following three core issues: (1) Whether this theoretical framework can effectively capture the aesthetic essence of modal invariance; (2) Compared with the existing advanced methods, does it have significant advantages in cross-modal aesthetic tasks? (3) Whether the representations learned have good interpretability.

4.1 Experimental Setup

Dataset

The experiment adopted the widely used cross-modal aesthetic benchmark dataset ArtMuse. This dataset contains over 10,000 high-quality samples, covering three art modalities: painting, music and poetry. Each sample is accompanied by aesthetic labels marked by experts (such as "romanticism", "abstract expression", etc.), and aesthetic scores based on multi-person evaluations are provided, offering a reliable basis for model training and evaluation.

Baseline model

To comprehensively evaluate the model performance, we selected the following three representative baselines for comparison:

●CLIP model: As a current advanced general cross-modal representation model, it performs exceptionally well in the task of natural image-text alignment.

●CNN-LSTM multimodal fusion model: It adopts a method that combines traditional feature extraction with sequence modeling, representing the classic multimodal learning paradigm.

● Simplified model: Including model variants that only use semantic pathways or emotional pathways to verify the necessity of the dual-path architecture proposed in this paper.

For the evaluation task, we conduct a systematic assessment of the model from the following three dimensions:

1. Zero-shot cross-modal aesthetic search: Use text descriptions as queries (such as "Search for works that embody a sense of sublimity"), conduct searches in image and music libraries, and take average search accuracy (mAP) as the core metric;

2. Cross-modal aesthetic classification: Train the classifier on a single modal (such as painting), and test it on other modalities (such as music), using accuracy as the evaluation metric to verify the modal invariance of aesthetic concepts;

3. Characterization quality analysis: By leveraging high-dimensional visualization techniques such as t-SNE, qualitatively analyze whether the structure of the characterization space is clustered based on aesthetic concepts rather than modal sources.

4.2 Result Analysis

The experimental results show that the cross-modal aesthetic unified representation model proposed in this paper demonstrates significant advantages in all evaluation tasks.

In the cross-modal aesthetic retrieval task, the retrieval accuracy rate of this model in the typical aesthetic category of "sublime" reached 0.82, which is 25 percentage points higher than that of the CLIP model, demonstrating an outstanding ability to capture abstract aesthetic concepts. This result confirms that the model can go beyond surface feature matching and achieve a deep understanding of the essence of aesthetics. In the cross-modal classification task, the model achieved an accuracy rate of 0.75 in the "painting → music" transfer task, significantly outperforming the CLIP model's 0.58. This performance advantage highlights that the learned representation has good modal invariance and can effectively support the transfer of aesthetic knowledge among different modalities. It is particularly worth noting that the model demonstrates an ability to learn general aesthetic laws, indicating that it has indeed captured the intrinsic aesthetic features that transcend specific modalities.

The visualization analysis of the representation space provides further support for the above conclusion. As shown in Figure 3, the representation space constructed by this method presents a clear clustering structure based on aesthetic concepts. Samples of different modalities of the same aesthetic concept are clustered in similar areas, while samples of different aesthetic concepts remain significantly separated. This spatial structure contrasts sharply with the modal information-dominated representation space of the CLIP model, intuitively demonstrating the effective extraction of aesthetic essential features by this method.

Based on the comprehensive quantitative results and qualitative analysis, the following conclusions can be drawn: Firstly, the aesthetic essential space hypothesis proposed in this paper can indeed effectively capture the aesthetic commonalities across modalities; Secondly, the dual-path mapping mechanism has successfully achieved the collaborative modeling of aesthetic semantics and emotions. Finally, the learning strategy guided by concept prototypes enables the model to discover meaningful aesthetic concept structures. These findings jointly verify the validity of the theoretical framework in this paper and provide new ideas and methods for the study of cross-modal aesthetic understanding.

5. Conclusions and Prospects

This paper systematically establishes a theoretical framework for the unified representation of cross-modal aesthetics. By innovatively proposing the "aesthetic essential space" hypothesis and the dual-path mapping mechanism, it successfully achieves a paradigm shift from traditional feature correlation to essential understanding. This framework not only provides a new theoretical perspective for solving the representation problems of cross-modal aesthetics, but more importantly, it constructs an aesthetic analysis system that is both computable and interpretable. The experimental results show that the model based on this framework can effectively capture the deep-seated aesthetic essential features while maintaining modal invariance, significantly improving the performance of cross-modal aesthetic tasks.

The theoretical innovation of this study is mainly reflected in three dimensions: First, the concept of aesthetic essential space with modal invariance is proposed, establishing a unified mathematical representation basis for aesthetic experiences of different modalities; Secondly, a semantic-emotion dual-path mapping mechanism was designed to achieve collaborative modeling of multi-dimensional features of aesthetic experience. Thirdly, it pioneered a contrastive learning method guided by concept prototypes, enabling the model to independently discover aesthetic concept structures with semantic significance. These theoretical breakthroughs not only promote the development of the discipline of computational aesthetics but also provide important methodological references for research in related fields.

Based on the current research results, future research will focus on the following three directions for in-depth exploration:

At the level of improving the theoretical system, we will be committed to building a more fine-grained and culturally adaptable prototype system of aesthetic concepts. The key research contents include: exploring the common aesthetic features in different cultural traditions and establishing cross-cultural aesthetic prototype representation methods; Analyze the semantic associations and hierarchical structures among prototypes to form a systematic aesthetic knowledge graph; Study the dynamic evolution laws of the prototype system and enhance the cultural adaptability and evolution ability of the model.

At the level of technological and methodological innovation, efforts will be focused on promoting the in-depth integration and development with generative AI. Specific research directions include: developing conditional generation models based on aesthetic representations to achieve controllable multimodal content generation; Research cross-modal aesthetic style transfer techniques and explore new paradigms for artistic creation; Build an interactive aesthetic creation system to promote creative work through human-machine collaboration. These studies will strongly promote the in-depth application of AIGC in the field of artistic creation.

At the level of expanding application scenarios, efforts will be made to build a dynamic and adaptive aesthetic computing framework. By introducing user feedback mechanisms and time series modeling techniques, the focus is on researching personalized optimization methods based on reinforcement learning, developing multimodal time series perception systems, enabling computational models to adapt to users' dynamic aesthetic preferences and achieving continuously evolving aesthetic understanding capabilities.

The theoretical value and practical significance of this research are equally important. At the theoretical level, it not only deepens the fundamental theoretical research of computational aesthetics, but also provides new methodological guidance for aesthetic computing in the context of human-machine-object integration. At the practical level, this framework has demonstrated broad application prospects in fields such as artistic creation, aesthetic education, and cultural heritage protection. With the continuous advancement of subsequent research, this theoretical framework is expected to lead artificial intelligence to achieve a leapfrog development from "perceiving beauty" to "understanding beauty" and then to "creating beauty" in the creative field, ultimately opening up new possibilities for human-machine collaborative creative work.

In the future, we will continue to deepen theoretical exploration, expand application boundaries, and focus on the cross-integration of aesthetic computing with disciplines such as brain science and cognitive science. We will explore the intrinsic connection between the neural mechanisms of aesthetic experience and computational models, and promote the construction of a more complete theoretical system of aesthetic intelligence.

References

[1] Radford A, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021.

[2] Arandjelovic R, Zisserman A. Look, listen and learn[C]//Proceedings of the IEEE international conference on computer vision. 2017.

[3] Li Yanzu. Introduction to Aesthetics [M]. Tsinghua University Press, 2006.

[4] Birkhoff G D. Aesthetic measure[M]. Harvard university press, 1933.