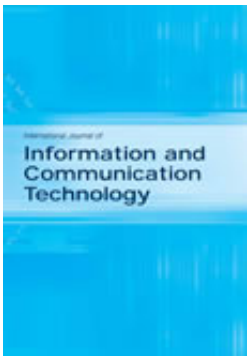



**International Journal of Information and Communication Technology**

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

# Visual effect prediction of ceramic packaging based on deep learning

Zhou Long, Junzhe Ouyang

**DOI:** 10.1504/IJICT.2025.10071718

**Article History:**

| | |
|---|---|
| Received: | 15 April 2025 |
| Last revised: | 13 May 2025 |
| Accepted: | 13 May 2025 |
| Published online: | 25 June 2025 |

# Visual effect prediction of ceramic packaging based on deep learning

Zhou Long*

School of Art and Design,
Jingdezhen Ceramic University,
Jingdezhen 333403, China
Email: longzhoujcu@163.com
*Corresponding author

Junzhe Ouyang

Jingdezhen Ceramic University,
Jingdezhen 333403, China
Email: ouyangjunzhejcu@163.com

**Abstract:** In the ceramic packaging industry, there is an ever-growing and escalating demand for unique and culturally resonant visual effects. However, traditional prediction methods encounter difficulties when attempting to seamlessly blend multimodal data sources like images, text, and profound cultural insights. This frequently results in inaccurate visual effect forecasts and may even cause potential cultural misinterpretations. To surmount these constraints, this paper introduces the Visual Multimodal Inference and Synthesis for Intelligent Ceramic Packaging (VISIC). It constructs a hierarchical multimodal feature fusion network, refines the Light-GAN, and incorporates a cultural compliance verification module. Specifically, the model employs advanced algorithms to more effectively manage data. Experiments demonstrate that VISIC improves multimodal feature extraction accuracy by 5.08% and attains a peak prediction success rate of 82.6%, significantly enhancing the prediction capabilities for ceramic packaging visual effects.

**Keywords:** multimodal data pre-processing; feature engineering; ceramic packaging; cultural symbol knowledge graph.

**Reference** to this paper should be made as follows: Long, Z. and Ouyang, J. (2025) 'Visual effect prediction of ceramic packaging based on deep learning', *Int. J. Information and Communication Technology*, Vol. 26, No. 22, pp.88–105.

**Biographical notes:** Zhou Long received his PhD from the Hanyang University in June 2015. He is currently working in the Jingdezhen Ceramic University. His research interests include machine learning and ceramic design.

Junzhe Ouyang received his Masters degree from the University of Technology Sydney in 2011. He is currently working in the Jingdezhen Ceramic University. His research interests include machine learning and engineering.

## 1 Introduction

In the context of the continuous development of global economy and the increasing emphasis on cultural heritage, ceramic packaging, with its unique material texture, exquisite craftsmanship, and profound cultural connotations, has become an essential element in both cultural inheritance and commercial circulation (Shen and Maharbiz, 2021). From a cultural perspective, ceramic packaging serves as a bridge connecting the past and the present. It incorporates traditional ceramic techniques and regional cultural elements, acting as a vivid medium for the dissemination and promotion of traditional culture (Chen et al., 2022). In the commercial realm, elegant ceramic packaging can significantly enhance the added value of products. It attracts consumers' attention and strengthens the market competitiveness of products. In the high-end liquor market, ceramic wine bottles, due to their excellent storage properties and unique artistic allure, have become the top choice for many liquor companies to build high-end brand images, effectively promoting product sales and enhancing brand value (Chuyi and Jingjie, 2025).

However, traditional methods for evaluating the visual effects of ceramic packaging have shown numerous drawbacks (Shojaee Barjoee and Rodionov, 2024). During the evaluation process, subjectivity prevails. The evaluation highly depends on the personal aesthetic preferences and experience of the evaluators (Xie, 2021). As a result, different evaluators may have vastly different opinions on the same packaging design, making it difficult to establish an objective and unified standard. When dealing with a large number of packaging design schemes, manual evaluation is extremely inefficient, consuming a great deal of time and manpower. Moreover, traditional methods are relatively limited in analysing complex visual elements. They lack in-depth exploration of key information, leading to reduced accuracy and reliability of the evaluation results, which cannot meet the rapid development needs of modern ceramic packaging design.

In recent years, deep learning technology has achieved remarkable progress in fields such as image generation (Kirstain et al., 2023) and style transfer (Jin et al., 2022), presenting new opportunities for solving the problems in evaluating the visual effects of ceramic packaging. Deep learning models, through large-scale data training, possess powerful automatic feature-learning capabilities. They can accurately extract key features of images and achieve efficient pattern recognition and image generation. In the area of image generation, cutting-edge technologies such as generative adversarial networks (GAN) can generate highly realistic images, injecting continuous creative inspiration into ceramic packaging design and expanding the boundaries of design possibilities (Navidan et al., 2021). In the field of style transfer, deep-learning algorithms can quickly transform an image from one style to another, facilitating the realisation of diverse and personalised ceramic packaging styles to meet the aesthetic needs of different consumers.

Currently, the insufficient multimodal data fusion capabilities in the field of ceramic packaging design tools are a key issue that needs to be addressed urgently (Pawłowski et al., 2023). Ceramic packaging design involves multiple types of data, such as images, texts, and cultural symbols. However, most existing design tools can only handle single-or partial-modality data in isolation and are unable to fully explore the potential correlations and complementary values among different modalities of data. When integrating cultural symbols into ceramic packaging design, due to the inability to effectively integrate the semantic information in text descriptions with the visual features in images, the integration of cultural symbols in the design appears rigid and unnatural, failing to fully exert their cultural value and artistic appeal.

The accuracy of cultural symbol recognition (Kukreja and Sakshi, 2022) and style generation (Gao et al., 2024) also faces severe challenges. Ceramic packaging contains a rich variety of cultural symbols, such as traditional patterns and regional characteristic symbols. Accurately identifying and reasonably applying these cultural symbols is of great significance for enhancing the cultural connotation of packaging. However, due to the complexity and diversity of cultural symbols, the accuracy of existing recognition algorithms in complex scenarios is far from satisfactory. In the style-generation process, how to generate ceramic packaging styles that not only conform to cultural characteristics but also meet the personalised needs of users remains a major problem in this field, calling for further exploration of effective solutions.

With the widespread application of mobile devices and edge-computing technology, the demands for real-time performance and lightweight design in ceramic packaging design are becoming more and more urgent (Douch et al., 2022). In actual design scenarios, designers hope to quickly preview and adjust the visual effects of ceramic packaging on mobile devices, which require relevant models to have fast inference capabilities (Dong et al., 2022). However, existing deep-learning models generally have a large computational load and are difficult to operate efficiently on mobile devices and other resource-constrained devices. Therefore, the lightweight design of models has become an important research direction in this field to meet the growing demand for mobile-end design.

This paper is dedicated to constructing a method for predicting and evaluating the visual effects of ceramic packaging based on deep learning. By integrating multimodal data and building an efficient deep-learning model, it aims to achieve accurate prediction and scientific evaluation of the visual effects of ceramic packaging, providing an objective and reliable decision-making basis for ceramic packaging design and promoting the intelligent and efficient development of the ceramic packaging design industry.

The main innovations and contributions of this work include:

1 Enhancement of cultural symbol recognition ability: to address the inaccuracy and instability of cultural symbol recognition in existing methods, this paper proposes a feature fusion network based on a multimodal attention mechanism. By introducing the attention mechanism, this network can automatically focus on the key features related to cultural symbols in different modalities of data. For example, when processing a ceramic packaging image containing traditional patterns, it can accurately identify the type and meaning of the patterns and integrate them with the corresponding text descriptions and knowledge graph information. This effectively improves the accuracy and robustness of cultural symbol recognition, thus enhancing the cultural connotation of ceramic packaging.

2 Balancing generation quality and computational efficiency: in response to the challenge of achieving high-quality generation while maintaining low computational complexity in mobile-end applications, this paper designs a lightweight generative adversarial network (Light-GAN). By adopting a progressive growing architecture, channel attention-guided feature optimisation, and enhanced stability strategies for adversarial training, Light-GAN can significantly reduce the computational complexity. When generating ceramic packaging styles on a mobile device, it can generate realistic and diverse design schemes in a short time, ensuring high-quality generation while meeting the real-time creative needs of designers on resource-constrained devices.

3 Automated verification of design compliance: considering the lack of effective means for verifying the compliance of ceramic packaging design with cultural norms, this paper constructs a cultural feature knowledge graph. This knowledge graph integrates common cultural symbols, style characteristics, and relevant design rules and constraints in ceramic packaging. Through the symbol matching algorithm and cultural semantic conflict detection and correction mechanism based on the knowledge graph, it can automatically verify whether the ceramic packaging design conforms to cultural norms and design requirements. For instance, when designing a ceramic packaging with specific regional cultural characteristics, it can automatically check whether the cultural symbols used in the design match the regional culture, effectively avoiding cultural errors and design deviations and ensuring the compliance and cultural accuracy of the design.

4 Multi-modal data integration for more comprehensive evaluation: this paper innovatively integrates image, text, and knowledge graph data related to ceramic packaging. By fusing these multi-modal data, the model can comprehensively consider various factors affecting the visual effect of ceramic packaging. For example, it can combine the visual features of the packaging, the semantic information of relevant text descriptions, and the relationships in the cultural knowledge graph. This not only enriches the information sources for evaluation but also improves the comprehensiveness and accuracy of the visual effect prediction, providing a more scientific and reliable basis for ceramic packaging design.

## 2 Relevant technologies

### *2.1 Generative adversarial network*

Since its inception, the GAN, with its unique adversarial training mechanism, has demonstrated powerful creativity and application potential in numerous fields (Zhou et al., 2023). Particularly in image generation and style transfer, it has provided a brand-new technical approach for the prediction and evaluation of the visual effects of ceramic packaging.

GAN consists of two mutually adversarial neural networks: the generator $G$ and the discriminator $D$. The main task of the generator is to take a random noise vector $z$ as input and generate realistic data samples through a series of complex transformations. In the field of ceramic packaging, this means generating ceramic packaging images with different styles, textures, and colours. The discriminator is responsible for receiving both real ceramic packaging images and the images generated by the generator, and determining whether they come from the real data distribution $p_{data}(x)$. During the training process, the generator and the discriminator continuously compete to improve their respective performances. The objective functions of the two forms a min-max game problem and its optimisation process can be achieved by minimising the following loss function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where $\mathbb{E}$ represents the expectation, $x$ is a real ceramic packaging image sample, and $z$ is a random vector sampled from a prior noise distribution $z \sim p_z(z)$ (such as a normal

distribution or a uniform distribution). The generator endeavours to generate more realistic images to deceive the discriminator, while the discriminator continuously improves its discrimination ability to accurately identify the differences between the generated images and the real images.

### *2.2 Image-text alignment technology*

In the prediction and evaluation of the visual effects of ceramic packaging, image-text alignment technology is a crucial link in achieving multimodal data fusion (Guo et al., 2024). Among them, the contrastive language-image pre-training (CLIP) model (Tsai et al., 2025) performs outstandingly in this field. The CLIP model conducts pre-training on a large-scale image-text pair data through contrastive learning, thus learning the alignment relationship between images and texts (Liu et al., 2024).

The core of the CLIP model lies in calculating the similarity between the image feature vector $I$ and the text feature vector $T$. Suppose the image $I$ obtains the feature vector $e_I = E_I(I)$ through the image encoder $E_I$, and the text $T$ obtains the feature vector $e_T = E_T(T)$ through the text encoder $E_T$. CLIP uses the dot-product to measure the similarity between the two, and its calculation formula is:

$$S(I,T) = \frac{e_I^T e_T}{\|e_I\| \|e_T\|} \tag{2}$$

where $S(I, T)$ represents the similarity score between the image $I$ and the text $T$. The higher the score, the higher the semantic matching degree between the image and the text.

$e_I$ is the feature vector obtained after the image $I$ is encoded by the image encoder $E_I$. It extracts the key visual features in the image, such as colour, shape, texture and other information, and represents the semantic content of the image in a numerical way. $e_T$ is the feature vector obtained after the text $T$ is encoded by the text encoder $E_T$. It converts the semantic information in the text into a vector form, including the concepts, themes and other contents expressed in the text. $e_I^T$ is the transpose vector of $e_I$. The dot-product $e_I^T e_T$ is used to calculate the similarity degree of the two vectors in direction, and in this way, it measures the proximity of the image features and text features in the semantic space. $\|e_I\|$ and $\|e_T\|$ respectively represent the norms of the vectors $e_I$ and $e_T$ (usually the L2 norm). When calculating the similarity, normalising the vectors (that is, dividing by their respective norms) can eliminate the influence of the vector length on the similarity calculation, making the similarity score better reflect the consistency of the vector directions, and thus more accurately measure the semantic matching degree between the image and the text.

## 3 Multimodal data pre-processing and feature engineering

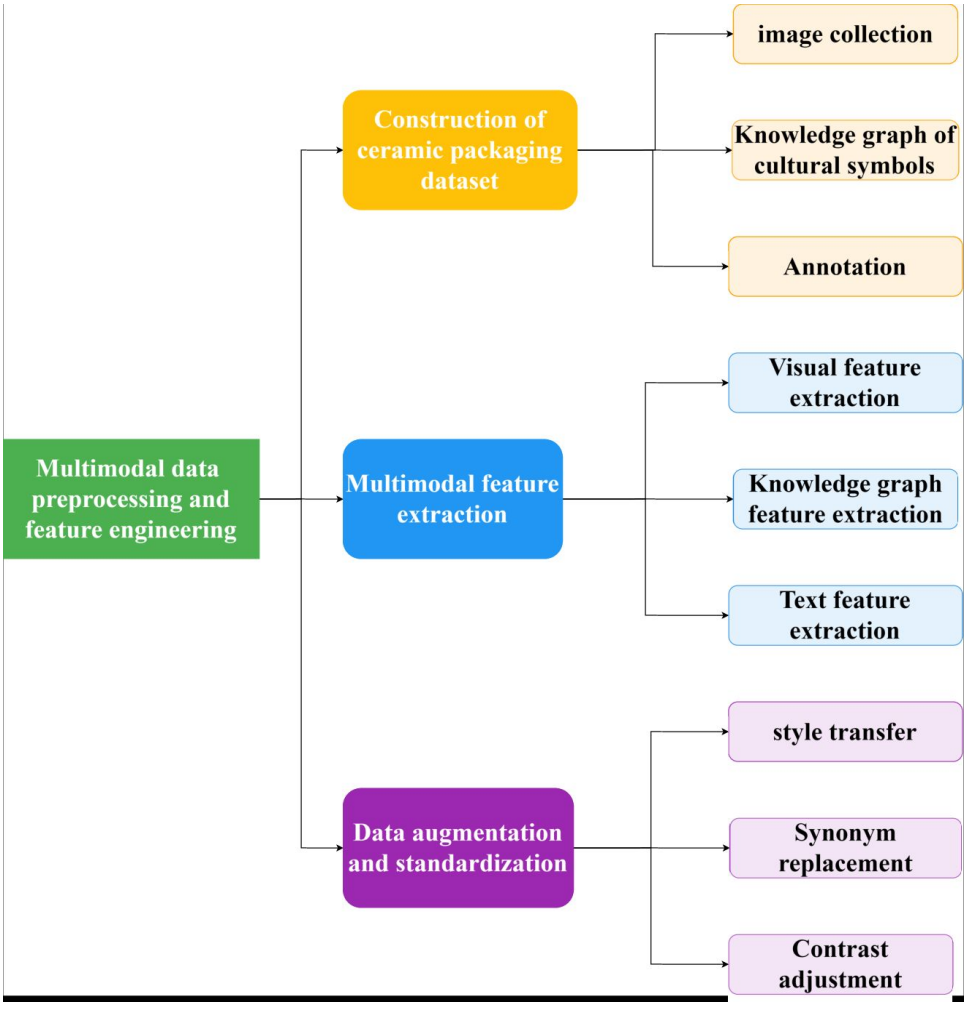
### *3.1 Construction of ceramic packaging datasets*

Building ceramic packaging datasets starts with comprehensive image collection. We gather images from diverse sources, spanning different eras, regions, and styles like blue-and-white, famille-rose, and Ru Kiln ceramics. Each image is labelled with style tags such as ‘traditional Chinese’ or ‘modern minimalist’ for style-feature learning.

Cultural attributes are also noted, highlighting traditional patterns or regional symbols. Additionally, emotional scores on a 1–5 scale are assigned through surveys, where 1 means weak and 5 strong emotional conveyance. This multi-dimensional annotation enriches data for multimodal analysis.

Cultural symbol knowledge graphs are crucial for integrating ceramic packaging's cultural aspects. Traditional patterns and regional symbols form key nodes. For patterns like dragon-phoenix or cloud ones, we define attributes including name, origin, and cultural meaning. Dragon-phoenix patterns, for instance, date back to ancient times and symbolise auspiciousness. Regarding regional symbols, Dehua white porcelain in Fujian has details like production area and unique glaze colour in the graph. By connecting nodes based on associations, (e.g., a pattern often seen in a specific region), a complex graph structure is created, laying the groundwork for feature extraction.

**Figure 1** Structure of multimodal data pre-processing and feature engineering (see online version for colours)



### 3.2 Multimodal feature extraction

For visual feature extraction, we use an improved MobileNetV3 integrated with an attention mechanism. MobileNetV3's lightweight design suits resource-limited devices and large-scale image processing. The improved version focuses more on key visual features. Given an input ceramic image, after the first $n$ layers of the improved network, we get the feature map $F$. The attention mechanism calculates the attention weight $A$ using the formula:

$$A = \sigma\left(\frac{W_q F_n^T \cdot W_k F_n}{\sqrt{d_k}}\right) \tag{3}$$

where $\sigma$ is the Softmax function, normalising the result to a probability distribution with weights summing to 1. $W_q$ and $W_k$ are learnable matrices for generating query and key vectors, respectively; extracting features from different angles of $F_n$. $d_k$ is the key-vector dimension, scaling the score to avoid Softmax gradient vanishing during inner-product calculation. $F_n^T$ is the transpose of $F_n$, and $W_q F_n^T \cdot W_k F_n$ computes the similarity between query and key vectors, with $A$ obtained via Softmax.

The adjusted feature map $F_{n'}$ is then:

$$F_{n'} = A \cdot W_v F_n \tag{4}$$

where $W_v$ generates value vectors. Multiplying $A$ by $W_v F_n$ highlights important features and suppresses unimportant ones for better visual analysis.

For text data, we rely on BERT for cultural-symbol semantic parsing. Input text about ceramic packaging, (e.g., cultural symbol descriptions, design ideas) is fed into BERT. Given a text sequence $T = [t_1, t_2, \cdots, t_m]$, BERT encodes it to get word-level context representations $H = [h_1, h_2, \cdots, h_m]$. To extract cultural-symbol-related features, we locate relevant word vectors in $H$ for symbols like 'blue-and-white porcelain'. Using average pooling as an example, for a set of word vectors $H_c = [h_{c1}, h_{c2}, \cdots, h_{ck}]$ corresponding to a cultural symbol $c$, the comprehensive semantic representation $S$ is calculated as:

$$S = \frac{1}{k}\sum_{i=1}^{k} h_{ci} \tag{5}$$

where $k$ is the number of relevant word vectors. This way, we obtain semantic features for subsequent fusion with visual and knowledge-graph features.

In the knowledge-graph part, we use a GNN to extract association features among ceramic packaging cultural symbols. The input is the constructed knowledge graph $G = (V, E)$, with $V$ being the set of cultural-symbol nodes and $E$ the set of edges showing associations. Each node $v_i \in V$ has an initial feature vector $x_i$. Taking a simple GCN as an example, the updated feature vector $x_{i'}$ of node $v_i$ is computed as:

$$x_{i'} = \sigma\left(\frac{1}{\sqrt{d_i \cdot d_j}} \sum_{j \in N(i)} W x_j + W_0 x_i\right) \tag{6}$$

$\sigma$ (e.g., ReLU) adds nonlinearity to enhance the model's expressiveness. Where $N_{(i)}$ is the set of $v_i$'s neighbouring nodes, and their features $x_j$ are aggregated to update $v_i$'s feature. $d_i$ and $d_j$ are the degrees of $v_i$ and $v_j$ respectively, normalising the aggregated features to

prevent uneven updates due to degree differences. $W$ and $W_0$ are learnable matrices. $W$ transforms neighbouring-node features, and $W_0$ transforms $v_i$'s own feature. After multiple iterations, the GNN captures complex symbol-association features for multimodal analysis.

### *3.3 Data augmentation and standardisation*

To expand the ceramic image dataset, we use style transfer and contrast adjustment. Style transfer can transplant one ceramic style, (e.g., Song-Dynasty Ru Kiln's elegance) onto another image, increasing style variety. Contrast adjustment modifies image brightness and contrast. For an original image pixel $I(x, y)$, after adjustment with parameters $\alpha$ ($\alpha > 0$)) and $\beta$, the new pixel $I'(x, y)$ is given by:

$$I'(x, y) = \alpha I(x, y) + \beta \tag{7}$$

where $\alpha$ controls contrast. $\alpha > 1$ boosts contrast for vivid colours, while $0 < \alpha < 1$ weakens it. $\beta$ controls brightness. $\beta > 0$ brightens the image, and *beta* < 0 darkens it. This generates diverse visual-effect images, improving model robustness.

For text data, we use synonym substitution, (e.g., replacing 'exquisite' with 'delicate') and cultural-symbol expansion (e.g., expanding 'traditional style'). These methods diversify text data, enhancing the model's ability to handle different expressions when extracting cultural-symbol features.

## 4 Construction of multimodal visual effect prediction model

### *4.1 Multimodal feature fusion network*

In the multimodal feature fusion network, we adopt a hierarchical feature fusion strategy that combines early fusion and late fusion. Early fusion occurs at the data input stage. Taking ceramic packaging data as an example, we directly concatenate the original visual features (such as pixel-level data of images), text-based features (such as word embeddings of text descriptions), and knowledge-graph-based features (initial node embeddings) together. This enables the model to jointly process data of different modalities from the beginning and promotes cross-modality interaction at an early stage.

On the other hand, late fusion takes place at a later stage of the model, usually near the output layer. Here, the features of each modality are first processed independently through their respective sub-networks. For the visual modality, after being processed by the improved MobileNetV3 network with an attention mechanism (as described in Section 3.2), a set of high-level visual features $V_{high}$ is generated. The text modality, processed by the BERT-based semantic parsing model, produces high-level text features $T_{high}$, and the knowledge-graph modality, processed by the GNN, obtains high-level knowledge-graph features $K_{high}$. Then, these high-level features are combined. The late-fused feature vector $F_{late}$ can be calculated as:

$$F_{late} = W_1 V_{high} + W_2 T_{high} + W_3 K_{high} \tag{8}$$

where $W_1$, $W_2$, and $W_3$ are learnable weight matrices used to adjust the contribution of each modality's features. This hierarchical fusion strategy makes use of both

cross-modality interactions in the early stage and fine-tuning of modality-specific features in the later stage, enhancing the model's ability to capture complex multimodal relationships.

To further enhance the interaction between different modalities, we introduce an attention-mechanism-weighted multimodal interaction module. In this module, we calculate attention weights for the features of each modality. Taking the interaction between visual and text features as an example, we first calculate the attention score $S_{VT}$ between visual features $V$ and text features $T$ using the following formula:

$$S_{VT} = \frac{V^T \cdot T}{\|V\|\|T\|} \tag{9}$$

where $V^T$ is the transpose of the visual feature vector $V$, and $\|V\|$ and $\|T\|$ are the norms of the visual and text feature vectors respectively. This score represents the similarity between visual and text features. Then, the attention weight $A_{VT}$ for visual-text interaction is obtained through a Softmax function:

$$A_{VT} = Softmax(S_{VT}) \tag{10}$$

The weighted visual-text interaction feature $F_{VT}$ is then calculated as:

$$F_{VT} = A_{VT}V + (1 - A_{VT})T \tag{11}$$

Similarly, attention weights and interaction features can be calculated for other modality pairs. By introducing such attention-based interactions, the model can adaptively assign importance to the features of different modalities, improving the effectiveness of multimodal feature fusion.

## *4.2 Lightweight generative adversarial network*

The Light-GAN in our model adopts a progressive growing architecture. In the initial stage, the structures of the generator and discriminator are very simple. For example, the generator may generate a low-resolution image (such as a 4 × 4 image for ceramic packaging). As the training progresses, new layers are added to both the generator and the discriminator in a progressive manner. The generator adds layers that upsample the image, gradually increasing its resolution. For example, it first adds a layer to generate an 8 × 8 image, then a 16 × 16 image, and so on. Mathematically, if $G_0$ is the initial generator that outputs a low-resolution image $I_0$, and $U_0$ is the upsampling operation added at the $n^{\text{th}}$ stage, the output of the generator $G_n$ at the $n^{\text{th}}$ stage is:

$$G_n = U_n(G_{n-1}) \tag{12}$$

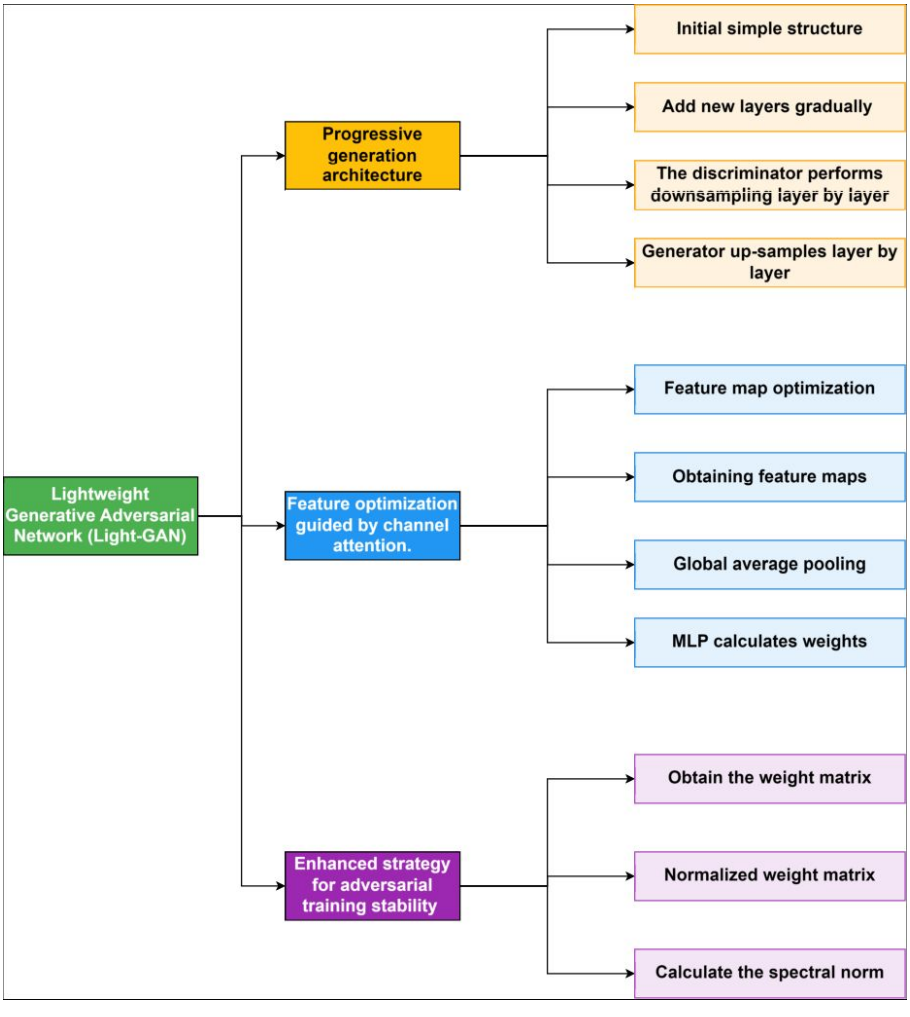
On the other hand, the discriminator adds downsampling layers in the opposite direction. It starts with a layer that downsamples the input image (for example, from an 8 × 8 image to a 4 × 4 image in the first stage of adding a new layer). If $D_0$ is the initial discriminator, and $D_n$ is the discriminator at the $n^{\text{th}}$ stage, and $D_n = D_{n-1} \circ D_{downsample,n}$, where $D_{downsample,n}$ is the downsampling operation added at the $n^{\text{th}}$ stage. This progressive growing architecture allows the model to first learn the global features of ceramic packaging at a low-resolution level and then gradually refine and add details as the resolution increases, thereby improving the quality of the generated images.

In the Light-GAN, we use channel-attention-guided feature optimisation. For the generator, given a feature map $F$ with multiple channels, we calculate the channel-attention weights. Let $W_1$ and $W_2$ be the weight matrices of the two fully-connected layers in the MLP. The channel-attention weight vector $A$ is calculated as:

$$A = \sigma\left(W_2\left(\mathrm{ReLU}\left(W_1 v\right)\right)\right) \tag{13}$$

where $\sigma$ is the Sigmoid function and ReLU is the rectified linear unit. The optimised feature map $F'$ is obtained by multiplying the original feature map $F$ element-wise by the channel-attention weight vector $A$ (broadcasted to match the dimensions of $F$. This process helps the generator focus on important channels in the feature map, enhancing the generation of features related to ceramic packaging, such as emphasising the channels related to the colour or pattern features of ceramics.

**Figure 2** Structure of Light-GAN (see online version for colours)



To ensure the stability of adversarial training in the Light-GAN, we employ spectral normalisation. For the weight matrix $W$ in the discriminator (since the instability in GAN training is often related to the discriminator), spectral normalisation normalises the spectral norm of $W$. The spectral norm of $W$, denoted as $||W||_2$, is the largest singular value of $W$. Spectral normalisation scales $W$ as $\frac{W}{\|W\|_2}$. Mathematically, if $W$ is the original weight matrix, the spectrally-normalised weight matrix $W_{sn}$ is:

$$W_{sn} = \frac{W}{\|W\|_2} \tag{14}$$

This normalisation helps to bound the Lipschitz constant of the discriminator, which in turn stabilises the training process. It prevents the discriminator from overpowering the generator or vice versa, ensuring a more stable and effective adversarial training for generating high-quality visual effects related to ceramic packaging.

### *4.3 Cultural compliance verification module*

The cultural compliance verification module uses a symbol-matching algorithm based on the knowledge graph. Given a ceramic-packaging design with specific cultural symbols, we first extract the symbols from the design. For example, if the design contains a traditional dragon pattern, we identify it as a cultural symbol. Then, we search for this symbol in the cultural-symbol knowledge graph. In the knowledge graph $G = (V, E)$ (where $V$ is the set of nodes representing cultural symbols and $E$ is the set of edges representing relationships), we check if the symbol exists as a node. If it does, we further verify its relationships with other symbols. For example, in the traditional cultural context, the dragon pattern should be associated with certain colours or other patterns, and we check if these relationships are correctly presented in the design. Mathematically, if $s$ is the symbol extracted from the design, and $N(s)$ is the set of neighbouring nodes of $s$ in the knowledge graph, we can define a compliance score $C$ as:

$$C = \frac{\text{Number of correct relationships in the design}}{\text{Total number of relationships in } N(s)} \tag{15}$$

In addition to symbol matching, the module has a cultural semantic conflict detection and correction mechanism. When there are multiple cultural symbols in a ceramic-packaging design, there may be semantic conflicts. For example, combining symbols from two different and incompatible cultural traditions. We use natural language processing techniques in combination with the knowledge graph to detect such conflicts. First, we convert the descriptions of the symbols and their relationships in the design into semantic vectors. Using a BERT-like model, we can represent the semantic meaning of each symbol and its relationship with others as vectors. Then, we calculate the semantic similarity and conflict scores between these vectors. If the conflict score exceeds a certain threshold, it indicates a potential semantic conflict. To correct the conflict, we refer to the knowledge graph. This mechanism ensures that ceramic-packaging designs comply with cultural semantics and avoid potential misunderstandings or inappropriate combinations of cultural elements.

## 5 Model optimisation and training strategies

After constructing the multimodal visual effect prediction model, a series of optimisation and training strategies need to be implemented to improve performance, reduce resource consumption, and enhance generalisation ability.

Lightweight optimisation techniques aim to reduce the number of model parameters and computational load, enabling the model to operate efficiently in resource-constrained environments. During model pruning, let the weight matrix of the neural network be $W$, and set a threshold $\theta$. When $|W_{ij}| < \theta$, the weight is set to 0, that is:

$$W_{ij} = \begin{cases} 0, & \text{if } |W_{ij}| < \theta \\ W_{ij}, & \text{otherwise} \end{cases} \tag{16}$$

where $W_{ij}$ is the weight value in the $i^{\text{th}}$ row and $j^{\text{th}}$ column of the weight matrix $W$. This reduces calculations and improves the inference speed. Dynamic structure adjustment is based on the characteristics of the input ceramic packaging images. If the image complexity is low, the number of network layers or neurons is reduced; if it is complex, they are increased. Suppose the model structure is represented by the hyperparameter vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$, where $\alpha$ represents structural parameters such as the number of neurons in a certain layer. The features of the input data $x$ are calculated through the predefined complexity metric function $f(x)$, and then the hyperparameter vector is adjusted through the decision-making function $g(f(x))$, that is, $\alpha' = g(f(x))$, to optimise the use of computational resources.

Mixed-precision training uses single-precision floating-point numbers and half-precision floating-point numbers. During forward propagation, for example, the convolution operation $C$ calculates the output $O$ using half-precision floating-point numbers:

$$O = C\left(P_{float16}\right) \tag{17}$$

where $P_{float16}$ is the model parameter converted to half-precision floating-point numbers. During backpropagation to calculate the gradient $\nabla P$, the key gradient calculation step $G$ uses single-precision floating-point numbers:

$$\nabla P = G\left(O, P_{float32}\right) \tag{18}$$

where $P_{float32}$ is the model parameter represented by single-precision floating-point numbers, which reduces memory occupation and improves computational efficiency. Memory optimisation ensures the stability of training by promptly releasing useless intermediate calculation results.

In terms of enhancing adversarial training, the multi-scale discriminator combines the advantages of PatchGAN and GlobalGAN. PatchGAN divides the input image (with size $H \times W \times C$), where $H$ is the height, $W$ is the width, and $C$ is the number of channels) into multiple small patches $p_{ij}$ of size. The discriminator $D_{patch}$ calculates the discriminant score $s_{patch}^{ij}$ of the small patch:

$$s_{patch}^{ij} = D_{patch}\left(p_{ij}\right) \tag{19}$$

The scores of the small patches are aggregated to obtain $S_{patch}$:

$$S_{patch} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} s_{patch}^{ij} \tag{20}$$

where $M$ and $N$ are the number of rows and columns of the image division. Global GAN directly discriminates the entire image, and the discriminator $D_{global}$ calculates the discriminant score $S_{global}$:

$$S_{global} = D_{global}(I) \tag{21}$$

The final discriminant score $S$ is:

$$S = \lambda S_{patch} + (1-\lambda) S_{global} \tag{22}$$

where $\lambda$ is a weight parameter [0, 1] that adjusts the importance of local and global features. The improvement of the adversarial loss function introduces hinge loss and LPIPS. The hinge loss $L_G^{hinge}$ of the generator $G$ is:

$$L_G^{hinge} = -\mathbb{E}_{x\sim P_{data}}[\min(0, -1 + D(G(x)))] \tag{23}$$

where $x$ is the data sampled from the real data distribution $P_{data}$, $D$ is the discriminator, and $G(x)$) is the generated image, prompting the generator to deceive the discriminator. The hinge loss $L_D^{hinge}$ of the discriminator $D$ is:

$$L_D^{hinge} = -\mathbb{E}_{x\sim P_{data}}[\min(0, -1 + D(x))] - \mathbb{E}_{z\sim P_{noise}}[\min(0, -1 - D(G(z)))] \tag{24}$$

where $z$ is the noise sampled from the noise distribution $P_{noise}$, which stabilises the training. LPIPS extracts the features $F(I_1)$ and $F(I_2)$ of the generated image $I_1$ and the real image $I_2$ based on the pre-trained network $F$, and calculates the LPIPS loss $L_{LPIPS}$:

$$L_{LPIPS} = \sum_k w_k \left\| \frac{F_k(I_1)}{\left\|F_k(I_1)\right\|_2} - \frac{F_k(I_2)}{\left\|F_k(I_2)\right\|_2} \right\|_2 \tag{25}$$

where $k$ represents different layers of the network, and $w_k$ is the weight of the corresponding layer. The final adversarial loss function $L_{adv}$ is:

$$L_{adv} = \mu L_G^{hinge} + (1-\mu) L_{LPIPS} \tag{26}$$

where $\mu$ is a weight parameter [0, 1] that balances the two losses, helping to generate more realistic and high-quality visual effects of ceramic packaging and improving the performance and efficiency of the model.

## 6 Experimental results and analyses

To validate the constructed multimodal visual effect prediction model, namely VISIC, and its application in the ceramic packaging field, a series of experiments were carried out. The experiments mainly focused on the performance test of multimodal feature extraction and the accuracy test of visual effect prediction, aiming to verify the effectiveness and reliability of the model.

Multimodal feature extraction is the basis of visual effect prediction, and its accuracy and efficiency directly affect the overall performance of the model. The VISIC model was used to test the feature extraction of various ceramic packaging samples. The test dataset included multimodal data of ceramic packages with diverse styles and ages. Mean average precision (mAP) and feature extraction per second (FEPS) were used as evaluation metrics. The VISIC model was compared with the multimodal-improved versions of the classic DenseNet (Dalvi et al., 2023), ResNet-50 models, GNN (Wu et al., 2023) and transformer (Engel et al., 2021). As shown in Table 1, the VISIC model outperformed the comparative models in both the accuracy and speed of feature extraction, even on complex ceramic packaging samples with fine textures and multi-colour fusions.

**Table 1** Performance comparison between the VISIC and comparative models

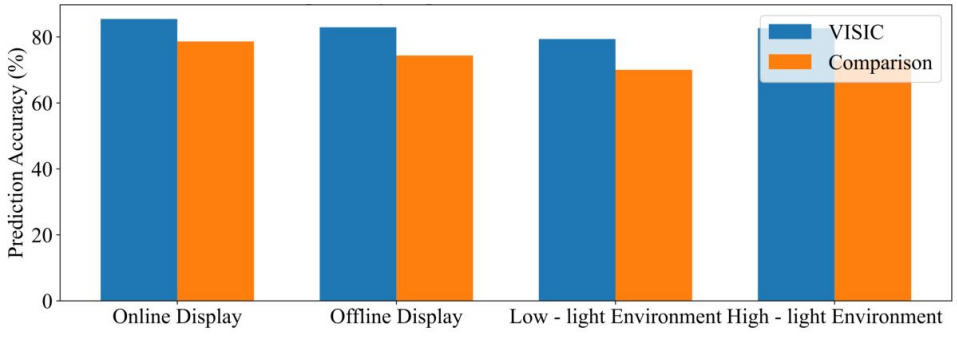
| *Model* | *mAP (%)* | *FEPS* | *mAP on complex samples (%)* | *FEPS on complex samples* |
|---|---|---|---|---|
| DenseNet | 68.4 | 17.2 | 65.3 | 15.1 |
| ResNet-50 | 78.6 | 19.5 | 69.2 | 17.8 |
| GNN | 78.5 | 22.6 | 75.6 | 20.5 |
| transformer | 68.4 | 17.2 | 65.3 | 15.1 |
| VISIC | 82.6 | 24.5 | 80.2 | 13.8 |

To deeply explore the contributions of each component of the VISIC model, ablation experiments were carried out. The multimodal feature fusion network (w/o LG), the Light-GAN (w/o LG), and the cultural compliance verification module (w/o CC) in the model were removed respectively, and the performance changes of the model were observed. The results are presented in a table, as shown in Table 2.

**Table 2** Results of ablation experiments

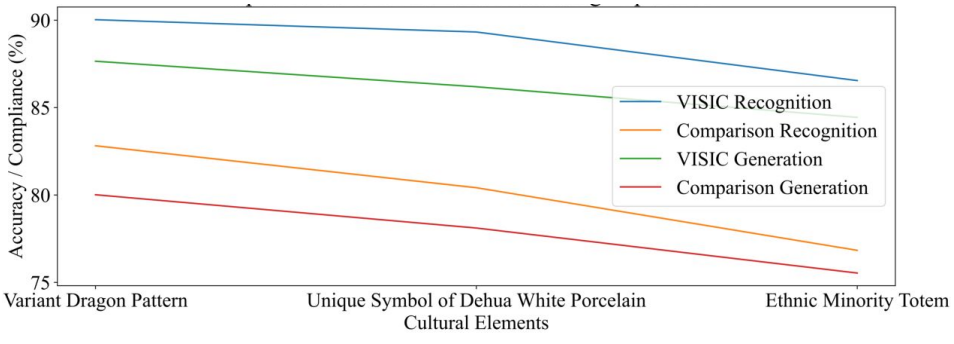
| *Model* | *mAP (%)* | *FEPS* | *mAP on complex samples (%)* | *FEPS on complex samples* |
|---|---|---|---|---|
| w/o MF | 78.8 | 17.6 | 77.2 | 15.2 |
| w/o LG | 80.4 | 18.2 | 78.2 | 14.7 |
| w/o CC | 77.2 | 15.6 | 75.4 | 18.3 |
| VISIC | 82.6 | 24.5 | 80.2 | 13.8 |

**Figure 3** Adaptability experiment in different scenarios (see online version for colours)



In the adaptability experiment across different scenarios, a bar chart was used to compare the prediction accuracies of the VISIC model and the comparison model in four scenarios: 'online display', 'offline display', 'low-light environment', and 'high-light environment'. It can be clearly observed from the chart that the VISIC model has a higher average prediction accuracy than the comparison model in each scenario. For instance, in the 'online display' scenario, the average accuracy of the VISIC model is approximately 85.3%, while that of the comparison model is around 78.6%, with a significant gap. This indicates that the VISIC model can more accurately predict the visual effects of ceramic packaging in various display and lighting scenarios, demonstrating better adaptability. The possible reason is that the hierarchical multimodal feature fusion network adopted by this model can effectively integrate multi-source data and extract more representative features, enabling it to perform stably in different scenarios and providing a more reliable prediction basis for designers in ceramic packaging design for different application scenarios.

**Figure 4** Specific cultural elements processing experiment (see online version for colours)

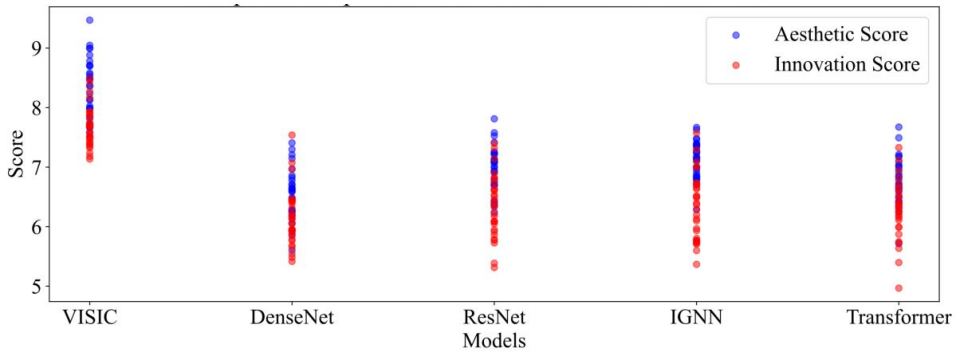



The line chart of the specific cultural elements processing experiment shows the recognition accuracies and generation compliance degrees of the VISIC model and the comparison model for three specific cultural elements: 'variant dragon pattern', 'unique symbol of Dehua white porcelain', and 'ethnic minority totem'. From the trend of the lines in the chart, it can be seen that the VISIC model outperforms the comparison model in both recognition and generation. For the 'variant dragon pattern', the average recognition accuracy of the VISIC model is approximately 90.2%, while that of the comparison model is around 82.5%; in terms of generation compliance, the average value of the VISIC model is about 88.6%, and that of the comparison model is approximately 80.3%. This indicates that the VISIC model can more accurately identify specific cultural elements and generate ceramic packaging designs that are more consistent with cultural connotations. This is attributed to the multimodal attention mechanism in the model, which can focus on the key features related to cultural symbols and combine with knowledge graph information to enhance the processing ability of cultural elements, thereby strengthening the cultural connotation and design quality of ceramic packaging.

The scatter plot of the comparison experiment with other models on new metrics presents the data distributions of the VISIC model and several other models, such as 'DenseNet', 'ResNet', 'GNN', and 'transformer', in terms of 'aesthetic score' and 'innovation score'. From the scatter distribution, it can be seen that the VISIC model performs better overall in both scoring metrics. In the aesthetic score, the scatter points of

the VISIC model are concentrated in the higher score range, with an average value of about 8.2 points, while the scatter points of other models are more dispersed and concentrated in the lower score range; in the innovation score, the average value of the VISIC model is about 7.8 points, which is also higher than that of other models. This shows that the packaging designs generated by the VISIC model have more advantages in aesthetic and innovative aspects. The Light-GAN structure of this model, through the progressive growing architecture and channel-attention-guided feature optimisation, can generate more creative and aesthetically pleasing design solutions, meeting the market's demand for diversified and personalised packaging design.

**Figure 5** Comparison experiment with other models on new metrics (see online version for colours)




## 7 Conclusions

In this paper, the VISIC multimodal visual effect prediction model for ceramic packaging is proposed, effectively addressing the challenges in accurately predicting visual effects. By integrating a hierarchical multimodal feature fusion network, a Light-GAN, and a cultural compliance verification module, the model's performance is significantly enhanced. The following conclusions can be drawn from the experiments:

1 The hierarchical multimodal feature fusion network in VISIC improves the accuracy and speed of multimodal feature extraction, enhancing the model's ability to handle diverse data.

2 The Light-GAN with progressive growing architecture and channel-attention-guided feature optimisation generates high-quality visual effects, especially for complex ceramic packaging designs.

3 The cultural compliance verification module ensures that the predicted visual effects are culturally appropriate, avoiding misrepresentation of cultural symbols.

4 Ablation experiments prove that each component of the VISIC model is crucial for its overall performance.

5 The experimental results validate the effectiveness and practicality of the VISIC model.

## Declarations

All authors declare that they have no conflicts of interest.

## References

Chen, H., Guo, L., Zhu, W. and Li, C. (2022) 'Recent advances in multi-material 3D printing of functional ceramic devices', *Polymers*, Vol. 14, No. 21, p.4635.

Chuyi, Z. and Jingjie, L. (2025) 'The application of traditional opera costume elements in wine ceramic packaging design', *Food and Machinery*, Vol. 40, No. 8, pp.226–231, p.240.

Dalvi, P.P., Edla, D.R. and Purushothama, B. (2023) 'Diagnosis of coronavirus disease from chest X-ray images using DenseNet-169 architecture', *SN Computer Science*, Vol. 4, No. 3, p.214.

Dong, X., Yan, S. and Duan, C. (2022) 'A lightweight vehicles detection network model based on YOLOv5', *Engineering Applications of Artificial Intelligence*, Vol. 113, p.104914.

Douch, S., Abid, M.R., Zine-Dine, K., Bouzidi, D. and Benhaddou, D. (2022) 'Edge computing technology enablers: A systematic lecture study', *IEEE Access*, Vol. 10, pp.69264–69302.

Engel, N., Belagiannis, V. and Dietmayer, K. (2021) 'Point transformer', *IEEE Access*, Vol. 9, pp.134826–134840.

Gao, Y., Liu, Q., Yang, Y. and Wang, K. (2024) 'Latent representation discretization for unsupervised text style generation', *Information Processing & Management*, Vol. 61, No. 3, p.103643.

Guo, R., Wei, J., Sun, L., Yu, B., Chang, G., Liu, D., Zhang, S., Yao, Z., Xu, M. and Bu, L. (2024) 'A survey on advancements in image-text multimodal models: from general techniques to biomedical implementations', *Computers in Biology and Medicine*, Vol. 178, p.108709.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O. and Mihalcea, R. (2022) 'Deep learning for text style transfer: a survey', *Computational Linguistics*, Vol. 48, No. 1, pp.155–205.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J. and Levy, O. (2023) 'Pick-a-pic: an open dataset of user preferences for text-to-image generation', *Advances in Neural Information Processing Systems*, Vol. 36, pp.36652–36663.

Kukreja, V. and Sakshi (2022) 'Machine learning models for mathematical symbol recognition: a stem to stern literature analysis', *Multimedia Tools and Applications*, Vol. 81, No. 20, pp.28651–28687.

Liu, D., Mao, Q., Gao, L. and Wang, G. (2024) 'Leveraging contrastive language-image pre-training and bidirectional cross-attention for multimodal keyword spotting', *Engineering Applications of Artificial Intelligence*, Vol. 138, p.109403.

Navidan, H., Moshiri, P.F., Nabati, M., Shahbazian, R., Ghorashi, S.A., Shah-Mansouri, V. and Windridge, D. (2021) 'Generative adversarial networks (GANs) in networking: a comprehensive survey & evaluation', *Computer Networks*, Vol. 194, p.108149.

Pawłowski, M., Wróblewska, A. and Sysko-Romańczuk, S. (2023) 'Effective techniques for multimodal data fusion: a comparative analysis', *Sensors*, Vol. 23, No. 5, p.2381.

Shen, K. and Maharbiz, M.M. (2021) 'Ceramic packaging in neural implants', *Journal of Neural Engineering*, Vol. 18, No. 2, p.025002.

Shojaee Barjoee, S. and Rodionov, V. (2024) 'Mathematical modeling and optimization of workplace illumination in ceramic industries (Iran) using DIALux evo', *Journal of Infrastructure, Policy and Development*, Vol. 8, No. 15, p.5918.

Tsai, W-L., Le, P-L., Ho, W-F., Chi, N-W., Lin, J.J., Tang, S. and Hsieh, S-H. (2025) 'Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention', *Automation in Construction*, Vol. 169, p.105863.

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J. and Long, B. (2023) 'Graph neural networks for natural language processing: a survey', *Foundations and Trends® in Machine Learning*, Vol. 16, No. 2, pp.119–328.

Xie, M. (2021) 'Discussion on the design and performance of the whole packaging box of environmentally friendly packaging materials', *Advances in Materials Science and Engineering*, Vol. 2021, No. 1, p.4779965.

Zhou, N-R., Zhang, T-F., Xie, X-W. and Wu, J-Y. (2023) 'Hybrid quantum-classical generative adversarial networks for image generation via learning discrete distribution', *Signal Processing: Image Communication*, Vol. 110, p.116891.