International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 1-17

https://doi.org/10.70917/ijcisim-2025-0263
Article

Research on the Application of Artificial Intelligence Algorithms in Image Content Creation in the Environment of Digital Transformation of Film Industry

Lili Xu *

Shandong Institute of Commerce and Technology, Jinan, Shandong, 250103, China; matongyue2023@163.com

Abstract: This paper proposes a new architecture for stylizing 3D meshes based on textual information based on the promotion of artificial intelligence technology for the digital transformation of the film industry. Combined with the object beam analysis of bi-directional reflection distribution function, the CLIP model is applied to the parameter optimization of a given frame to realize the 3D mesh stylization for realism rendering. FID, LPIPS and CLIP score metrics are used to analyze the performance of the model generated based on CLIP model-driven 3D mesh stylization. And the image content generated by the model is scored. Among them, the image content generated by this method scores 36.85 and 39.14 in accuracy and authenticity, respectively.The similarity value between text and image tested on MS-COCO dataset reaches 0.312, and the result measured by this paper with VQA evaluation metrics on CelebA-HQ dataset is 0.894.Combining the performances of all the metrics, this paper proposes the CLIP-based model-driven 3D network stylized generation model can be used for movie image content creation and can be further promoted.

Keywords: CLIP model; bi-directional reflection distribution function; 3D grid; FID index; movie and television content creation

1. Introduction

In November 2021, the “14th Five-Year Plan for China's Film Development” issued by the State Film Bureau clearly mentioned the important goal of significantly enhancing the scientific and technological capacity of films, and promoting the digital transformation of films. The digital transformation of the film will be a long-term systematic project, which not only requires a comprehensive and integrated layout from a strategic height, but also requires the collaboration of the whole chain of the film and television industry [1-3]. Compact digital infrastructure construction, all-round integration and docking industry resources, and gradually establish a film digital production ecosystem centered on digital assets and a film interactive consumer experience ecosystem centered on data elements, and interlinked to form a cycle of interaction, so as to build up a new pattern of film digitalization industry driven by new quality productivity [4-8]. Promote the development of film digital creation tools, virtual production technology, etc., and promote the widespread application of new technologies such as 4K, high frame rate, giant screen projection, laser projection, etc., in digital film projection technology, to comprehensively improve the audio-visual quality of digital films, enhance the audience's sense of presence and sense of visual impact, so that different viewers can get a better moviegoing experience, and thus promote the digital film in the traditional media and audio-visual new media competition [9-13].

In 2024, the total box office of Chinese movies was 42.502 billion yuan, down about 22.6% from 2023. In 2024, the total box office of North American movies was $8.57 billion, down 3.8% from 2023.In 2024, the total box office of North American movies was $8.57 billion, down 3.8% from 2023. However, the overall shrinkage of the global box office does not mean a decline in the level of film

production and the number of production, but rather the structural contradiction between film content and market demand, as well as the deep convergence of digital media has prompted the audiovisual content market to carry out its own structural reorganization [14-16]. The digital transformation of film is not only an external drive to promote technological iteration, improve production efficiency, expand artistic imagination and enhance innovation, but also a systematic strategic thinking and panoramic top-level design to reshape the film industry [17-19]. The deep integration of artificial intelligence (AI), cloud computing, Internet of Things (IoT), blockchain, 5G, big data, real-time rendering engine and other technologies is facilitating the upgrading of the film industry, optimizing the creation of film content, and realizing the two-way interaction between film content and audience needs.

The application of AI in the field of film and television covers from script creation, post-production to movie viewing interaction. In terms of script generation and optimization, [20] proposed a natural language processing-based movie script generation system, which automatically generates scripts or stories based on existing movie storylines by optimally integrating deep learning and heuristic algorithms. Literature [21] proposes an automated movie generation model called "MovieFactory", which generates detailed movie scripts through user text-based ChatGPT, and introduces visual generation and audio retrieval to generate audio and video content that matches the storyline, which minimizes the movie production process and cost. Literature [22] uses natural language processing to extract text world elements to generate a base script under the "script-to-graph" approach, introduces reinforcement learning algorithms for evaluation, generates personalized scripts by optimizing different plot directions, and promotes screenwriters to innovate narrative and character development. Literature [23] integrated and combined a multi-stage genetic optimization model and ChatGPT to adjust the parameters of character motivation, plot coherence, and dialogue fluency in movie or TV drama scripts, which effectively optimized the quality of the scripts and gained higher emotional resonance. Literature [24] used the BERT model to classify movie scripts emotionally in the context of AI and introduced an emotion optimization algorithm to optimize the scripts, which improved the emotional quality of the original scripts.

In terms of script visualization preview and virtual filming, literature [25] designed a pre-visualization tool based on a movie database, which provides visual elements for screenwriters to refer to by previewing the text and dialogues of the script and promotes the screenwriters to target changes in the content of the script. Literature [26] developed an AI-based Cine Vision platform for visual pre-visualization of scripts, which provides visual references for directors and cinematographers to provide character customization, directing style simulation, and dynamic lighting setups during the film production process. Literature [27] used AI-based virtual reality technology to optimize the 3D animation sub-scene design, machine learning and neural network to generate personalized animation, which improved the efficiency and quality of the production and realized the virtual shooting of the film. Based on the movie script, literature [28] integrated multiple AI tools to create various functional modules such as extracting scripted scenes, generating visual layouts, deploying digital assets in the 3D engine, and real-time camera planning to realize automated and immersive virtual production. In addition, literature [29] mentions that AI technology combined with virtual reality and augmented reality can visualize and plan movie content and combine sentiment analysis and personalized recommendations to enhance user interaction experience and movie consumption.

In terms of image effects and real-time rendering, literature [30] used Unity 3D engine for 3D animation simulation under the background of AI and big data technology, and designed an animated character motion trajectory planning strategy based on particle swarm optimization algorithm, which was inputted into the engine for testing, and the visual effect was real and memorable. Literature [31] used deep learning algorithms to generate real-time rendering of moving image effects, obtaining higher quality and faster results than simulated rendering, in terms of quality the image is more detailed performance and texture realism. Literature [32] proposed an AI-based film and television special effects extraction algorithm, through the extraction of special effects parameters, the introduction of fuzzy neural algorithms to process the parameters and output to the MATLAB software to calculate the corresponding value of the special effects, the extracted special effects are clearer, free from the impact of the screen size, and provide data support for the production of film and television special effects. Literature [33] constructs an improved 3D convolutional neural network for multi-frame film and television special effects needs, by capturing the key spatio-temporal features in the movie image and performing ultra-high resolution processing, it improves the realism and quality of the texture and visual narrative, and enhances the sense of immersion in the film and television content.

In terms of restoration and enhancement of image content, literature [34] combines bilateral filtering and convolutional neural networks for filtering and feature extraction of old movie images, and introduces a deep learning network based on hopping connections to optimize the loss function to achieve high-quality restoration of old movie speckle noise. Literature [35] used generative adversarial

network and self-attention mechanism to improve the detail and color restoration of female hero images in red films, optimized the video augmentation quality, and used temporal convolutional network to ensure the consistency and smoothness of video frames. Literature [36] introduced an AI-based coding damage repair processing framework in conjunction with super-resolution techniques for video for improving the quality of multilevel degraded video and solving the problem of video during transmission due to broadband and storage limitations.

In terms of intelligent editing affecting content and user interaction, literature [37] shared an intelligent editing algorithm for film and television with smooth shot switching, climax capture, and action plot recognition, using convolutional neural networks to extract image features in film and television, long and short-term memory networks to extract temporal dependencies in video, an attention mechanism to fuse multimodal information of audio and text, and self-supervised learning and transfer learning techniques to meet specific editing needs. Literature [38] provides an AI generation framework for movie trailer editing, which consists of text-video generation network and video-music generation network modules to acquire multimodal features and explore the associations with a transformer model to generate high-quality trailers. Literature [39] proposes that a gesture recognition algorithm based on convolutional neural network-support vector machine is used as a technical support for human-computer interaction in virtual reality film and animation to promote in-depth interaction between the audience and the animation work, and to improve the audience experience. Literature [40] developed a 3D film and television animation production cloud management system, using AI assisted tools to achieve efficient, high-quality animation production, and combined with real-time preview and feedback based on the user to optimize the details of the animation content and the sense of realism, and form a deeper interaction with the audience.

While AI brings new life to film and television content creation, issues such as security, copyright, and ethics arise. Literature [41] concludes that AI-generated content does not lose its protectiveness during human participation and supervision, but models similar to those with mimicry styles and repetitive input data may infringe, and unauthorized third-party use may also exist. Literature [42] emphasizes that digital movie assets may cause cybersecurity issues due to piracy and extortion and suggests the use of cryptography, digital rights management, blockchain, and AI for cybersecurity maintenance. Literature [43] constructs a pioneering blockchain trusted digital copyright management system that is efficient, secure, and transparent, tracks digital assets in full cycle with the help of blockchain technology, and verifies the authenticity of assets by combining with digital watermarking technology. Literature [44] points out that automated content identification/auditing tools supported by AI technology can assist in the identification of video content infringement in online platforms, assist platforms in notifying and removing infringing content, reduce legal risks, and protect image copyrights.

In conclusion, AI uses multimodal information and audience information to generate movie video clips, and optimizes special effects and real-time rendering of traditionally shot movie videos, shortening the production cycle while improving the quality of production.AI combines with blockchain, digital watermarking technology, etc., to protect digital movie copyrights, but it also requires that all of mankind comply with relevant intellectual property rights laws and regulations, and work together to safeguard the interests of the creator, and to promote the The continuous deepening of the digital transformation and upgrading of the film industry.

This paper analyzes the current ecological development of the film digital industry and explains the promotion of artificial intelligence technology to the film industry from different aspects. It briefly describes the new trend of movie image narrative in the era of artificial intelligence, and explains the interactive narrative of digital images from multiple perspectives. Propose the object's bidirectional reflection distribution function BRDF and ambient lighting, use the pre-trained text-image macromodel CLIP as a supervisor, optimize the BRDF material, geometric, and ambient light parameters, and form a three-dimensional mesh stylization based on the semantic information driven by CLIP to optimize the intelligent creation of image content for the TV industry. Image-to-image and text-to-image generation tasks are carried out to evaluate the performance of the model in terms of quality, diversity, and semantic similarity by conducting comparative experiments, and to evaluate the value of the model for the creation of movie image content.

2. Intelligent image development for the film digitization industry

2.1. Film Digitalization Industry Ecology

(1) Overview of the film digital industry ecosystem

Film digital industry ecology refers to an industrial ecosystem formed by the combination of digital technology and the film industry, including digital cinema, digital production platform, digital distribution, digital projection, digital marketing platform, digital derivatives and other contents. These

contents are interrelated and promote each other, constituting a complete film digital industry ecosystem.

(2) The Development of Film Digital Industry Ecosystem

With the development of digital technology, movies have made great progress in production, distribution, promotion, screening and copyright management, forming a complete movie digital industry chain.

First: digital technology is applied to all aspects from script writing to post-production, such as special effects production, photography, editing, post-production and soundtrack. The application of virtual reality (VR) and augmented reality (AR) technologies in movie production is gradually increasing.

Second: Digital projection needs to be based on digital cinema and digital projection technology.

Third: The application of digital technology makes the movie ticketing system easier to operate.

Fourth: With the development of digital technology, new types of movie projection modes such as cloud movies and online theaters are emerging. The digital production platform provides more resources and support for the production of movies, and improves the efficiency and quality of movie production.

Fifth: Digital technology has brought richer marketing ideas and communication channels for movies.

Sixth: Movie production companies have begun to pay more attention to the analysis of data, focusing on audience preferences, market trends, box office and other information.

Seventh: The internationalization of the film industry.

2.2. Artificial Intelligence Advances Digital Transformation Changes in the Film Industry

Artificial intelligence technology can boost the innovation and development of the film industry from three aspects.

First: artificial intelligence technology helps the efficient production of film and television works.

The development of generative artificial intelligence technology makes the production of dialog-based content more convenient. Film and TV drama creators can input their own script direction, plot development needs, character image construction and other dialog content, and with the assistance of the AIGC application, they can quickly come up with film and TV scripts and character shaping images for reference, which greatly improves the efficiency of film and TV work content creation.

In addition, with the assistance of cross-modal content generation technology, the application of AI technology throughout the entire film industry process will become increasingly common. From scriptwriting, stage design, special effects production, virtual rehearsals, digital twin full-scene modeling, expression capture, automatic editing, to the formulation of marketing strategies, all will become emerging digital dividends for the film industry in the era of artificial intelligence. Film enterprises can also further reduce production costs through the continuous update and in-depth application of technology.

Second: Artificial intelligence technology improves the quality of movie production management.

The core production material of generative artificial intelligence is massive big data, and through network big data mining, collation and analysis and computational feedback, artificial intelligence can realize high-efficiency management of the whole chain of film production.

Third: AI technology promotes the optimization and reengineering of film industry processes.

With the powerful digital technology capability of instantaneous data mining, market monitoring and distributed computing of AI technology, the process of film industry is constantly optimized and reengineered, which can maximize the synergy between the content and the market, and make the film content creation and market operation a two-way fit.

2.3. Interactive Image Narratives in the Age of Artificial Intelligence

2.3.1. Digital Image Interactive Narrative

(1) Interactive Images

The interactive image is a form of image that allows the viewer to interact with it. This interaction is reflected in the mutual reaction and influence between the viewer and the content of the image.

(2) Video-Mediated Narrative

Video narratives have been different from other narratives since their inception.

Due to the characteristics of the subject matter, video narrative requires the author to complete the narrative function and realize the transmission of consciousness within a limited time, so the narrative approach adopted by the author is particularly important. Although the history of the development of video narrative is still relatively short compared with other types of narrative, video producers have begun to realize the powerful vitality of video as a narrative text. The most important feature of video

narrative is that video sound is directly visual and audible for the viewer, which is more explicit and specific than other narrative modes. Elements of the image, such as depth of field, light, color, compositional changes, etc., can also make the narrative produce different characteristics.

(3) Immersive Interactive Image Narrative

An interactive video narrative work is a multimedia hypertext experience that synthesizes text, image and sound. It integrates the qualities of different art forms while blurring the boundaries between various media.

This paper defines the different forms of interactive video narratives based on the four modes of interaction: reading, audiovisual, experiencing and participating, which range from shallow to deep.

Immersive Interactive Video Narrative is an advanced narrative that integrates interactive technology, video art and immersive experience. It transforms the viewer from a passive viewer to a participant capable of interacting with the story through Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR) technologies. Together with sensor technologies such as motion capture and eye tracking, the audience's body movements and gaze can become part of the interaction.

2.3.2. Interactive Image Narrative Based on Intelligent Technology

(1) Narrative convergence of image media

With the passage of time and the advancement of technology, media has experienced an evolution from early written symbols, to printing, radio and television, to the present new media era.

In the information age, where technologies such as artificial intelligence, bioinformatics, and virtual reality have converged and brought about innovations, the media has gone beyond the role of merely serving as an extension of human perception and has become an indispensable part of human life. This transformation further highlights the interactivity and integration between human beings and the media, which is no longer a one-way relationship but interdependent and mutually reinforcing.

(2) Multi-dimensionality of image media narrative

Image design has transcended the traditional geometric spatial dimensions, such as one-dimensional to four-dimensional limitations, creating a broader creative field. Five-dimensional and multidimensional spatial interaction design goes even further, providing unlimited possibilities for narrative, allowing content to unfold in multiple dimensions, creating unprecedented narrative depth and breadth.

From the perspective of visual symbols, image design does not only rely on conventional fonts, graphics and colors, but continues to explore diverse image forms and multi-dimensional design strategies. The advancement of technology has gradually shifted the image from static to dynamic, bringing more abundant expression tools for design.

(3) Narrative Orientation of Image Media

Narrative orientation of image media focuses on how images shape and guide the audience's perception and interpretation.

(4) Narrative Interactivity of Image Media

In the digital age, video narratives have become more diversified, and their structures are like trees or webs, intertwined and complex.

3. Text-to-image content technology

3.1. Text processing network-pre-training model CLIP

CLIP is a pre-trained model based on large-scale natural language processing and computer vision, and a 400million image-text dataset was constructed for training at the beginning of the training period.CLIP is a multimodal model that is capable of understanding both natural language and visual content. It learns to map linguistic and visual features to each other by training on a large number of image-text pairs, which enables it to reason about the relationship between images and text. This gives CLIP powerful performance on a wide range of tasks, including image classification, natural language reasoning, text categorization, and visual quizzing. A contrast learning technique is used during training, which enables the model to learn a representation that is more sensitive to features between similar image and text pairs. Also, CLIP uses a multi-task learning strategy, which enables the model to learn more generalized feature representations by training on multiple tasks [45-46].

CLIP consists of two models: a text encoder and an image encoder, where the text encoder is used to extract textual features, using the Text Transformer model commonly used in natural language processing. During the pre-training process, the text encoder learns to map different words and phrases into corresponding vector representations. The image encoder is used to extract the features of the image, and in this paper the Visual Transformer (Vi T) network is used as the image encoder.

CLIP needs to input both text features and image features during training, which makes the model learn

the function of graphic matching, that is, when text features are input, the model will match to the corresponding image by calculating the cosine similarity.

The training process is shown below:

Comparative matching learning is performed on text features T extracted by the text encoder and image features I extracted by the image encoder. For a training batch containing N graphic matching pairs, the N text features and N image features are required to be matched one by one, respectively.The CLIP model matches the N text features and N image features and predicts the similarity of the $2N$ possible matching pairs. Here the cosine similarity of text features and image features is calculated based on the text tokens in the text input and image tokens in the image input.The training objective of CLIP is to maximize the similarity of N positive samples while minimizing the similarity of $N^2 - N$ negative samples.

The cosine similarity (CS) is calculated as:

$$\cos\_similarity = \frac{A \cdot B}{\|A\| * \|B\|} \tag{1}$$

where A and B denote two vectors, respectively. - denotes the vector dot product, * denotes the vector fork product, $\|A\|$ denotes the mode of vector A.

During the training process and using symmetric cross-entropy (SCE), the loss is mainly targeted at the problem of noisy labels, preventing the network from fitting to the wrong labels and thus optimizing the model. Symmetric cross-entropy is a loss function for classification problems that can be effectively applied to unbalanced datasets or datasets with category bias. The formula for symmetric cross entropy is:

$$SCE(p,q) = -\frac{1}{N}\sum_{i=1}^{N}\left(\alpha y_i \log(p_i) + (1-\alpha)(1-y_i)\log(1-p_i)\right) \tag{2}$$

where p is the predicted output of the model, q is the distribution of true labels, y_i denotes the true label of the i th sample, p_i denotes the i th sample, N is the number of samples, and α is a weight coefficient to control the weights of the different categories.

3.2. Bidirectional reflection distribution function

The surface of an object receives radiation from a light source and emits reflected light externally, and for an observer, every facet of the target's surface is a source of radiation.

In a camera imaging system, the camera has an aperture center of O, an optical axis of OO_1, a surface point S_0 at a distance of f_p from the camera, and an image plane at a distance of f_p from the camera. The surface point S_0 of the target object projects the point S_p on the image. S_p is the intersection of S_pO and the image plane, similar to a point source of radiation. If the radiance of the reflected light from an object is L_r, the angle α between the surface normal and the line connecting the point on the surface of the object to the center of the camera. Image of a pixel surface element area dS_0 in the image plane projected area of dS_p, dS_0 in this direction projected area of $dS_0 \cos\alpha$, received by the camera surface element dS_p, assuming conservation of energy. Let θ_r be the angle between the plane normal and the line connecting the incident light aperture nodes. The stereo angle formed by S_p and the lens is ω, and the radiant flux $d\Phi_r$ emitted by dS_0 to dS_p is:

$$d\Phi_r = dS_0 \int_{\Omega_r} L_r d\Omega_r \tag{3}$$

where Ω_r is the projected stereo angle corresponding to the center of the aperture, the stereo angle formed by the center point of the aperture O and dS_0 is of the same magnitude as the stereo angle

formed by it and dS_p . $\theta_{r'}$ is the angle between the surface normal and the line connecting the point to the center of the aperture, so:

$$\frac{dS_0 \cos\theta_{r'}}{f_0^2} = \frac{dA_p \cos\alpha}{f_p^2} \tag{4}$$

Thus the image gray level is:

$$E_p = \left(\frac{f_0}{f_p}\right)^2 \cos\alpha \int_{\omega_r} L_r \frac{\cos\theta_r}{\cos\theta_{r'}} d\omega_r \tag{5}$$

If the diameter of the aperture d is assumed to be small relative to its distance from the surface of the object, $\theta_r \approx \theta_{r'}$, the L_r of the Lambertian body can be considered to be fixed, and the stereo angle occupied by the center of the aperture is:

$$\omega_r = \frac{dA_0}{dl^2} = \frac{\frac{\pi}{4} d^2 \cos\alpha}{f_0^2} \cos^2\alpha \tag{6}$$

Thus equation (5) reduces to:

$$\begin{aligned} E_p &= \left(\frac{f_0}{f_p}\right)^2 \cos\alpha \int_{\omega_r} L_r \frac{\cos\theta_r}{\cos\theta_{r'}} d\omega_r \\ &= \left(\frac{f_0}{f_p}\right)^2 \cos\alpha \cdot L_r \cdot \frac{\frac{\pi}{4} d^2 \cos^3\alpha}{f_0^2} \\ &= L_r \cdot \frac{\pi}{4} \left(\frac{d}{f_p}\right)^2 \cos^4\alpha \end{aligned} \tag{7}$$

This proves that the image irradiance is proportional to the scene irradiance and proportional to the square of the camera lens diameter. The expression may deviate in other imaging systems, but still satisfies $E_p \propto L_r$.

After understanding the radiometry and the camera imaging system, the Lambertian bidirectional reflection distribution function can be better rationalized. The Bidirectional Reflection Distribution Function (BRDF) is used to describe the reflection coefficients between an incident and an outgoing beam [47].The BRDF, denoted by the symbol f_r, describes the governing transmission properties of an object's surface as a result of reflections, and is a functional model for modeling the reflection properties of nonsmooth surfaces. It indicates how bright a surface is when it is illuminated by light from one direction and viewed from another direction.BRDF shows the reflection characteristics of a target object to a light source, different target objects have different reflection function models, and the ideal diffuse reflection is called Lambertian reflection.

A surface-specific coordinate system is established, with the Z axis in the direction of the local normal along the local normal of the surface, and XOY as the local tangent plane. θ denotes the zenith angle, which is the angle between the light ray and the surface normal, and ϕ denotes the azimuth angle, which is the angle between the projection of the light ray in the XOY plane and the X axis. (θ_i, ϕ_i) is the angle of incidence and (θ_r, ϕ_r) is the angle of reflection. E_i is the incident irradiance and L_r is the reflected irradiance. In the (θ_i, ϕ_i) direction, the radiation flux $d\Phi_0$ received on dS_0 is:

$$d\Phi_0 = L_i \cos\theta_i \cdot d\omega_i \cdot dS_0 = L_i \cdot d\Omega_i \cdot dS_0 = dE_i \cdot dS_0 \tag{8}$$

Similarly, for the reflection angle (θ_r, ϕ_r), the radiative flux emitted by dS_0 is:

$$d\Phi_r = L_r \cos\theta_r \cdot d\omega_r \cdot dS_0 = L_r \cdot d\Omega_r \cdot dS_0 \tag{9}$$

BRDF is the ratio of the observed reflected irradiance dL_r to the irradiance dE_i in the incident direction:

$$f_r\left(\theta_i,\phi_i;\theta_r,\phi_r\right) = \frac{dL_r\left(\theta_i,\phi_i;\theta_r,\phi_r,E_i\right)}{dE_i\left(\theta_i,\phi_i\right)} \tag{10}$$

The surface of a Lambertian body reflects all incident light without absorption or transmission, the radiance outgoing M is equal to the incident irradiance E_i, and the reflected irradiance L_r of the reflected light is the same in all directions. According to the relevant definitions of radiometrics, there are:

$$\begin{aligned} M &= \int_{\Omega_r} L_r d\Omega_r = \int_{\Omega_r} L_r \cos\theta d\omega \\ &= \int_{-\pi}^{\pi}\int_{0}^{\frac{\pi}{2}} L_r \cos\theta \sin\theta d\theta d\phi = L_r\pi \end{aligned} \tag{11}$$

This results in a BRDF of Lambertian bodies:

$$f_r = \frac{L_r}{E_i} = \frac{L_r}{M} = \frac{1}{\pi} \tag{12}$$

The BRDF of the Lambertian body determines the nature and characteristics of its surface diffuse reflection.

3.3. CLIP model-driven 3D mesh stylization

This paper proposes an end-to-end architecture for stylizing a given 3D mesh based on input text without any 3D data training. Specifically, this chapter uses the pre-trained text-image macromodel CLIP as a supervisor to iteratively optimize the BRDF material, geometry (normal offset) and ambient light parameters in the framework to achieve realistic and robust 3D mesh stylization.

(1) Intersection of camera ray and mesh

For a given input 3D mesh M and target text, this paper requires a series of operations to prepare the data and environment for image rendering and compositing. Detailed step-by-step descriptions are given below:

(a) Scale the mesh M into the unit sphere. Perform a scaling operation on the input 3D mesh M so that it lies inside a unit sphere.

(b) Determine the anchored viewpoint on the sphere.

(c) Randomly sample other camera positions.

(d) Calculate camera rays.

(e) Compute the intersection of the ray with the grid.

(f) Calculate the normal at the intersection point.

(2) 3D mesh surface reflection modeling

In order to generate 3D mesh M stylization with photo-level realism, this paper uses three components to model the appearance: environment mapping, normal mapping, and SVBRDF.

In this paper, we choose to use a spherical Gaussian function to efficiently approximate the closed form rendering equations. Spherical Gaussian functions have two extremely useful properties when computing integrals, and these properties ensure ease and efficiency of computation.

(a) The product of two spherical Gaussian functions is still a spherical Gaussian function. Specifically, if there are two spherical Gaussian functions, denoted $G\left(v;\mu_1,\lambda_1,a_1\right)$ and $G\left(v;\mu_2,\lambda_2,a_2\right)$, and their product can be expressed as:

$$\begin{aligned} &G\left(v;\mu_1,\lambda_1,a_1\right)G\left(v;\mu_2,\lambda_2,a_2\right) \\ &= G\left(v;\frac{\mu_m}{\left\|\mu_m\right\|},\left(\lambda_1+\lambda_2\right)\left\|\mu_m\right\|,a_1a_2e^{\lambda_m\left(\left\|\mu_m\right\|-1\right)}\right) \end{aligned} \tag{13}$$

where $\lambda_m = \lambda_1 + \lambda_2$ and $\mu_m = \dfrac{\lambda_1\mu_1 + \lambda_2\mu_2}{\lambda_1 + \lambda_2}$.

(b) The integral of a single spherical Gaussian function has a closed form solution:

$$\int_{\Omega} G(v;\mu,\lambda,a)dv = 2\pi\frac{a}{\lambda}\left(1-e^{-2\lambda}\right) \tag{14}$$

In this paper, multiple spherical Gaussian functions are used to represent the environment mapping $L_i(\omega_i)$. The environment mapping is an important part of describing the lighting conditions, which can affect the appearance and realism of the rendering results. The environment mapping is represented as follows:

$$L_i(\omega_i) = \sum_{k=0}^{M} G\left(\omega_i;\mu_k,\lambda_k,a_k\right) \tag{15}$$

The ambient light energy $E(L)$ is the sum of the contributions of the individual spherical Gaussian functions in the ambient mapping, and is used to represent the total energy of the ambient light. It can be calculated by the following equation:

$$E(L) = \sum_{k=0}^{M}\frac{2\pi a_k}{\lambda_k\left(1-e^{-2\lambda_k}\right)} \tag{16}$$

To further provide more geometric detail to the mesh at rendering time, for each point x_p and normal $n_p \in \left\{\left(1,\theta_p,\varphi_p\right)\middle|\theta_p \in (0,2\pi),\varphi_p \in (0,\pi)\right\}$, a normal offset is estimated in this paper to obtain an estimated normal $\hat{n}_p \in \left\{\left(1,\hat{\theta}_p,\hat{\varphi}_p\right)\middle|\hat{\theta}_p \in (0,2\pi),\hat{\varphi}_p \in (0,\pi)\right\}$:

$$(\hat{\theta}_p,\hat{\varphi}_p) = (\theta_p + \Delta\theta,\varphi_p + \Delta\varphi) = \Pi\left(\beta(x_p,\theta_p,\varphi_p);\gamma\right) \tag{17}$$

$$(\Delta\theta,\Delta\varphi) = \Delta\Pi\left(\beta(x_p,\theta_p,\varphi_p);\gamma\right) \tag{18}$$

where $\Delta\Pi$ is a multilayer perceptron used to estimate the offset of the normal.

In order to better simulate the reflective properties of the 3D model surface and produce more realistic stylized rendering results, this paper uses SVBRDF to model the surface reflective properties. In particular, SVBRDF consists of a diffuse reflection term and a specular reflection term. The diffuse reflection term can be expressed as:

$$f_d(x_p) = \frac{\Phi_1(\beta(x_p);\xi_1)}{\pi} \tag{19}$$

The SVBRDF function for specular reflection can be written as:

$$f_s(v,\omega_i,x_p) = \sum_{j=0}^{N} G\left(h;\hat{n}_p,\frac{\lambda_j}{4h\cdot v},\mathrm{M}\ {}_pa_j\right) \tag{20}$$

where h is a semivector representing the average direction of the incident direction ω_i and the observation direction v.

Ultimately, the complete SVBRDF function can be represented as a combination of the diffuse and specular reflection terms:

$$f_r(v,\omega_i,x_p) = f_d(x_p) + f_s(v,\omega_i,x_p) \tag{21}$$

This function describes the reflective properties of a surface under different lighting and viewing directions, including the effects of diffuse and specular reflections. By adding these two components, this paper is able to obtain more comprehensive and realistic rendering results.

(3) Rendering

From the SVBRDF obtained in the previous section, after offsetting the normal and ambient lighting, the color of the ray Rp can be rendered according to the following steps.

First, the color of each pixel needs to be determined. In addition, in order to obtain richer geometric details, this paper predicts the normal offset for each intersection point x_p, which results in an estimated normal $\hat{n}_p$. Next, a deep learning model was used to predict the parameters of diffuse and specular reflections. Then, the estimated normals $\hat{n}_p$, environment mapping parameters $\{\mu_k, \lambda_k, a_k\}$, and SVBRDF $\Phi\left(x_p; \xi\right)$ and view direction v_p, the color of the ray Rp can be calculated. Repeating the above steps for each pixel will allow them to eventually be summarized together to form the complete rendered image.

(4) Text-based correlation loss function

The optimization process in this paper is supervised by a pre-trained CLIP model, which in turn is a multimodal model designed to handle the correlation between images and text.

In each iteration, this paper first randomly samples a set of camera positions. Then, when proceeding to the data enhancement stage, this paper will randomly crop part of the regions of the image I and resize them to the size of (224,224), thus obtaining the enhanced image I_a. At the same time, the input text cues are also encoded by the text encoder of CLIP to obtain its corresponding latent encoding $\mathrm{L}_t \in \mathbb{R}^{512}$.

Finally, this paper adopts cosine similarity as the overall optimization objective to compute the similarity between potential codes L_i and L_t:

$$Loss(\mathrm{L}_i, \mathrm{L}_t) = -\frac{\mathrm{L}_i \cdot \mathrm{L}_t}{|\mathrm{L}_i|_2 |\mathrm{L}_t|_2} \tag{22}$$

The goal of this loss function is to maximize the similarity between the image encoding and the text encoding, thus ensuring that the generated image is similar to the given textual cues in the multimodal space. In this way, this paper is able to guide the image generation in a textual descriptive manner, leading to a more semantically aware rendering.

4. Content generation for digitized three-dimensional images of films

4.1. CLIP-based model-driven 3D stylized generation

(1) Image-to-Image Generation Results Comparison

In this paper, we evaluate the performance of the proposed CLIP model-driven 3D web-based stylized generation model in terms of quality, diversity and semantic similarity through comparative experiments.

In the experiments, this paper uses 3000 images from the test set in Multi-Modal-CelebA-HQ as the source images, and each image generates a target image for the FID metrics. For the 3D network stylization generation task, this paper is compared with five methods, IC-GAN, LAFITE, CLIP2LATENT, SDG and ILVR. The results of FID metrics calculation for the generation task are shown in Fig. 1, and the FID metrics results of SDG algorithm and ILVR algorithm are similar. As can be seen from the figure, the results of the FID metrics of this paper's model in multiple generation tasks range from 13.9 to 16.4, which is significantly lower than those of the other compared methods. In terms of image quality assessment, this paper's method performs best in terms of FID metrics. This indicates that the images generated by this paper's method are closer to real images and have higher visual quality.

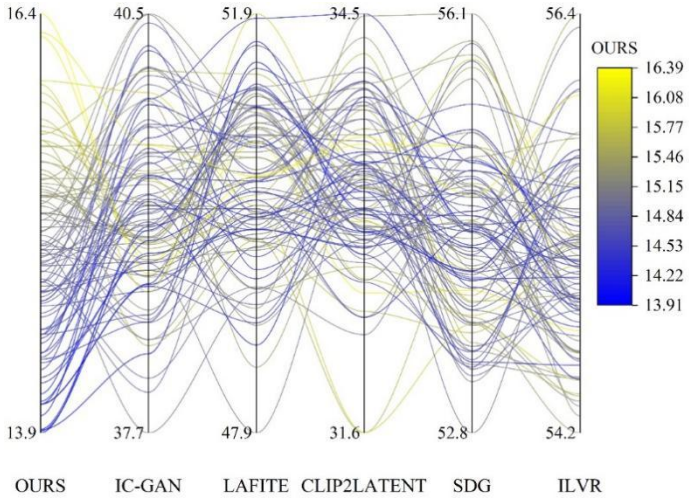

Figure 1. The fid index of the generated task is calculated.

In this paper, 100 face images are randomly selected for LPIPS and CLIP score calculation. Specifically, in calculating the LPIPS, this paper generates two target images for each input source image, and then calculates the LPIPS scores for 100 pairs of target images. The LPIPS scores of each method are shown in Fig. 2.

The LPIPS scores of this paper's method are in the range of 0.2 to 0.9 and are concentrated around 0.5. The LPIPS scores of LAFITE method are in the range of 0.0 to 1.2, and the maximum value of the LPIPS scores is higher than that of this paper's method. In terms of LPIPS metrics, this paper's method performs well and still achieves a high score despite not being optimal.

It should be noted that the LAFITE method performs best on the LPIPS metrics though. However, through further analysis, it is found that this may be due to the poor quality of the images generated by LAFITE itself, which leads to inaccuracies in the calculation of LPSPS values. Therefore, the method in this paper is still able to provide a certain degree of diversity while maintaining image quality, providing users with a wider range of options and application scenarios.

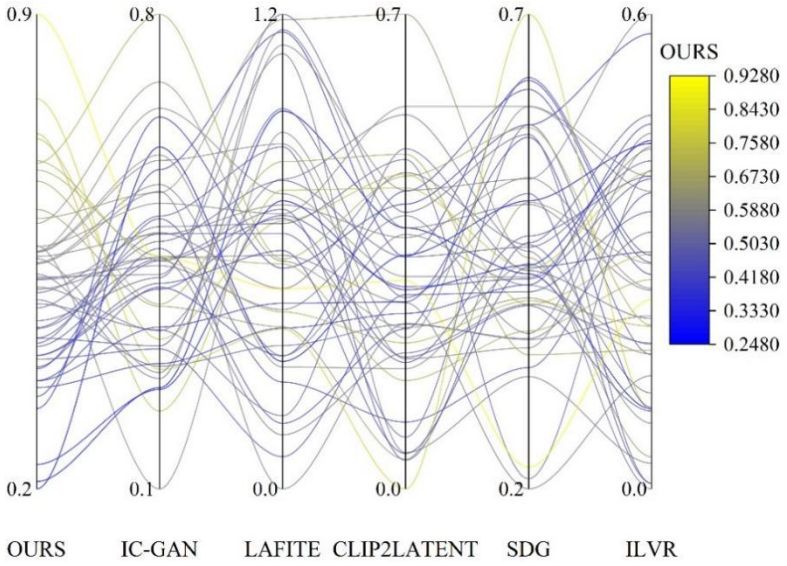

Figure 2. LPIPS scores for all parties.

Since CLIP can be batch processed, in calculating CLIP score, this paper generates 32 reference images for each source image. Then the best CLIP score is selected from these 32 candidate samples.

The CLIP score of each method is shown in Fig. 3.

The CLIP score value in the aspect of this paper is closer to the optimal value in multiple image generation task measurements and the mean value of CLIP score score is 0.914. The mean value of CLIP score score of other methods are IC-GAN (0.724), LAFITE (0.822), CLIP2LATENT (0.892), SDG (0.767) and ILVR (0.797). In terms of semantic similarity, this paper's method is significantly higher

than the results of the other five methods, which indicates that this paper's method performs best in terms of semantic similarity.

The experimental results show that this paper's method exhibits very good performance in all three metrics. In practical applications, this paper's method has the potential to generate high-quality, diverse and semantically consistent images, providing a powerful solution for image generation tasks.

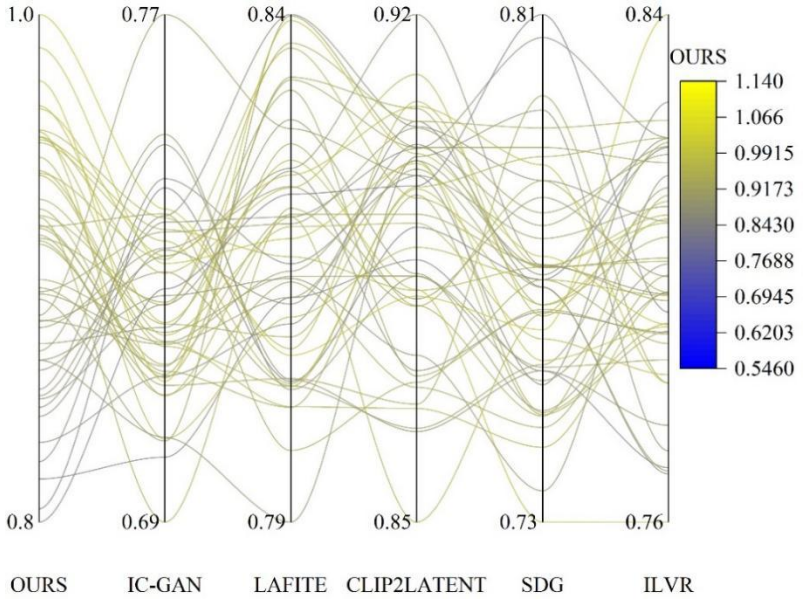

Figure 3. CLIP score of all parties.

(2) Comparison of text-to-image generation results

For the CLIP semantics-driven 3D web-based stylized generation model, this paper compares with five recent methods CLIP2LATENT, LAFITE, SDG, TediGAN and DALLE-2.

The quantitative comparison of text-to-image generation tasks is shown in Fig. 4.

The LPIPS scores of this paper's method and the comparison methods are 0.725, 0.568, 0.471, 0.601, 0.566, and 0.801, respectively.The LPIPS scores of this paper's method are lower than those of the DALLE-2 method but higher than the other comparison methods. Although the output of DALLE-2 performs better, it requires huge computational resources.

In terms of CLIP score metric data, the CLIP score of this paper's method is 0.359, which is lower than the TediGAN method's 0.365, with a difference of only 0.006.

CLIP2LATENT is able to generate images that basically fit the input text description better, but the images look a little unnatural.LAFITE and SDG generate images that are richer than CLIP2LATENT, but the image quality looks average.

Considering the performance of the two metrics, LPIPS and CLIP score, the method in this paper performs best overall.

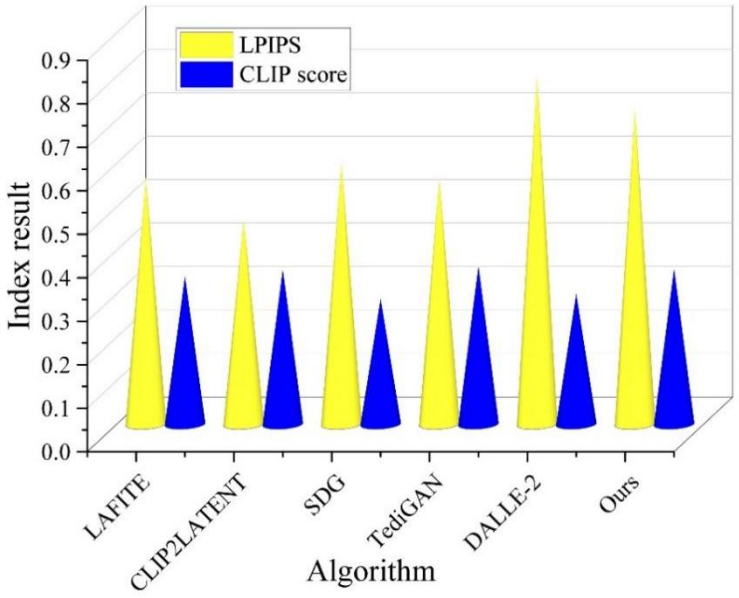

Figure 4. Text to image generation task quantitative comparison.

4.2. Comparative experiment on text-guided image content creation

In order to assess the accuracy and authenticity of the content of the images generated by this method, this paper presents a human subjective assessment of the generated results from the comparative methods and a random collection of 32 images that were manipulated on the basis of 32 textual cues.Fifty participants with diverse backgrounds were asked to vote for the best result based on three equally important principles.First, they were supposed to choose the result whose semantics corresponded most closely to the textual cues. Second, they should choose the outcome that best preserves human identity. Third, they should choose the image that is the most visually realistic. Users made objective evaluations of the accuracy and authenticity of the images in terms of their conformity to the semantics of the text, and user support was recorded in a quantitative manner.

The experiment recorded the subjects' judgments and compared them with the results generated by other current methods, and the results of the users' objective evaluations are shown in Fig. 5.

The methods compared are classical GAN-based methods, including AttnGAN, ControlGAN, DFGAN, DM-GAN and TediGAN, which all have excellent generation performance.

However, the evaluation comparison reveals that the image content generated by this method scores 36.85 and 39.14 in accuracy and authenticity, respectively, and this method surpasses the previous methods. Thus, the effectiveness of the text-to-image content generation method in this paper is verified.

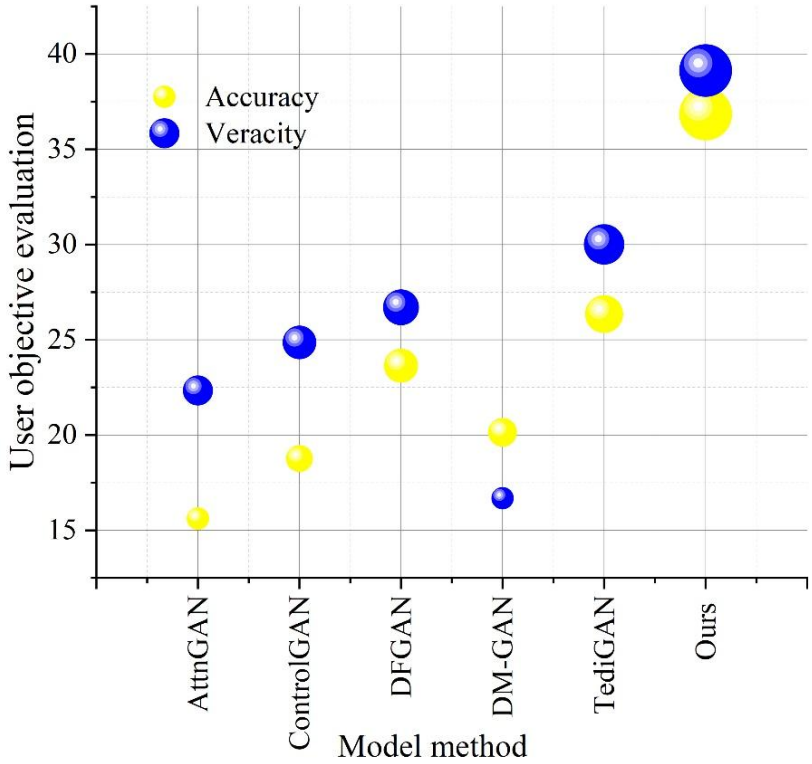

Figure 5. User objective evaluation results.

Earlier in this paper, the generative performance of the model of this method and the quality of the generated image content were subjectively evaluated by manual subject assessment.

The similarity values of the present method on different datasets are shown in Table 1, which demonstrates the results of the objective evaluation, including the text-image similarity on the MS-COCO dataset and the VQA scores on the CelebA-HQ dataset.

The similarity value between text and image tested on the MS-COCO dataset reaches 0.312, which is significantly higher than the other compared methods. This indicates that this method has the ability to generate images that match the semantics of the text with better performance on complex data. On CelebA-HQ dataset this paper compared other methods with VQA evaluation metrics test, the results show that this method is slightly higher than ManiGAN method, which indicates that this method generates image content that conforms to the semantic information and the user's ideal content.

Table 1. The similarity of the method of the different data sets.

Method	MS-COCO	CelebA-HQ
	$(\tilde{x},t)$	VQA score
AttnGAN	0.256	0.653
ControlGAN	0.148	0.414
DFGAN	0.301	0.659
DM-GAN	0.098	0.722
TediGAN	0.207	0.654
SISGAN	0.103	0.521
TAGAN	0.119	0.648
ManiGAN	0.113	0.759
Ours	0.312	0.894

5. Conclusion

In this paper, the visual language pre-training model CLIP is utilized to drive movie image content creation. By iteratively optimizing the BRDF material, geometry and ambient light parameters in the framework from the CLIP model, the 3D mesh stylization optimization in the digital transformation of movies is achieved.

(1) In the image-to-image generation task, the FID metric value of the CLIP model-driven 3D network-based stylized generation model proposed in this paper is lower than that of the five comparative

methods, namely IC-GAN, LAFITE, CLIP2LATENT, SDG and ILVR. The maximum value of FID index of this method is 16.4, which is lower than the maximum value of ILVR method by 40. In terms of image quality assessment, the method of this paper has the best performance. Considering the three metrics data together, the method of this paper is superior in solving the image-to-image generation task.

(2) The LPIPS and CLIP score metrics data of this paper's method are 0.725 and 0.359 in text-to-image generation results, respectively. In terms of LPIPS score, it is second only to the large-scale model DALLE-2, which indicates that this paper's method has a significant advantage in terms of diversity. Considering the performance of both LPIPS and CLIP score, this paper's method has the best performance in the whole.

(3) The quantitative evaluation results of text information-guided generation of movie image content show that this paper's method can generate image content that better fits the text semantics on complex data, and the CLIP model-driven generation of movie image content is more accurate and authentic.

REFERENCES

[1] Tsiavos, V., & Kitsios, F. (2025). The digital transformation of the film industry: How Artificial Intelligence is changing the seventh art. Telecommunications Policy, 103021.

[2] Weinberg, C. B., Otten, C., Orbach, B., McKenzie, J., Gil, R., Chisholm, D. C., & Basuroy, S. (2021). Technological change and managerial challenges in the movie theater industry. Journal of Cultural Economics, 45(2), 239-262.

[3] Schulz, A., Eder, A., Tiberius, V., Solorio, S. C., Fabro, M., & Brehmer, N. (2021). The digitalization of motion picture production and its value chain implications. Journalism and Media, 2(3), 397-416.

[4] Odoh, A., Ogungbe, F., & Olagunju, A. T. (2024). Digital Transformation in sub-Saharan Africa's Film Industry. Journal of African Films and Diaspora Studies, 7(4), 165.

[5] Maria Sadlowska, K., Sonja Karlsson, P., & Caldwell Brown, S. (2019). Independent cinema in the digital age: is digital transformation the only way to survival?. Economic and Business Review, 21(3), 5.

[6] Paksiutov, G. D. (2021). Transformation of the global film industry: Prospects for Asian countries. Transformation.

[7] Mbura, I. (2022). Effects of Digitalization on the Three-tier Structure of Tanzania's Film Industry. Umma: The Journal of Contemporary Literature and Creative Art, 9(1), 140-163.

[8] Barile, D., Secundo, G., & Magnusson, M. (2025). Exploring the digital innovation ecosystem from the perspective of platform-based startups: a case study in the film industry. European Journal of Innovation Management, 1-24.

[9] Jurgess, T. (2017). DIGITAL CINEMA AND ECSTATIC TECHNOLOGY: frame rates, shutter speeds, and the optimization of cinematic movement. Angelaki, 22(4), 3-17.

[10] Lopez, M., Kearney, G., & Hofstädter, K. (2022). Seeing films through sound: Sound design, spatial audio, and accessibility for visually impaired audiences. British Journal of Visual Impairment, 40(2), 117-144.

[11] Wang, J., Mou, X., & Wen, C. (2022). Measurement of image quality for large-screen displays based on audience viewing: Overview of standardization and preliminary study on laser projection displays and LED displays. Journal of the Society for Information Display, 30(6), 523-530.

[12] Kotlińska, M. (2024). The Influence of Digital Transformation on the Evolution of the Audiovisual Industry. European Research Studies Journal, 27(S2), 429-443.

[13] Yang, X., & Khoo, O. (2024). Projecting China to the world: Cinity, high frame rate cinema and the future of Chinese screening technology. Asian Cinema, 35(1-2), 63-79.

[14] Pustu-Iren, K., Sittel, J., Mauer, R., Bulgakowa, O., & Ewerth, R. (2020). Automated Visual Content Analysis for Film Studies: Current Status and Challenges. DHQ: Digital Humanities Quarterly, 14(4).

[15] Giannatou, E., Campagnolo, G. M., Franklin, M., Stewart, J. K., & Williams, R. (2019). Revolution postponed? Tracing the development and limitations of open content filmmaking. Information, Communication & Society, 22(12), 1789-1809.

[16] Indriasari, R., Hadiningrat, K. S. S., & Wardani, F. P. (2025). AN OPPORTUNITIES AND CHALLENGES OF THE INDONESIAN FILM INDUSTRY BASED ON DIGITAL PROMOTION TRANSFORMATION. JIPOWER: Journal of Intellectual Power, 2(1), 73-91.

[17] Liu, H., & Zheng, S. (2025). Adapting to Crisis: Social Media Behavior and Strategic Shifts of Chinese Cinema Chains During COVID-19. SAGE Open, 15(3), 21582440251363720.

[18] Puspitasari, L., Bajari, A., Hidayat, D. R., & Cho, S. K. (2024). Technological innovation and social construction of Makassar film industry's production and distribution. Jurnal Kajian Komunikasi, 12(2), 254-266.

[19] Awobamise, A. O. (2018). Michael Curtin, Jennifer Holt & Kevin Sanson (Eds): Distribution Revolution. Conversations about the Digital Future of Film and Television. MedieKultur: Journal of media and communication research, 34(65), 158-161.
[20] Dharaniya, R., Indumathi, J., & Kaliraj, V. (2023). A design of movie script generation based on natural language processing by optimized ensemble deep learning with heuristic algorithm. Data & Knowledge Engineering, 146, 102150.
[21] Zhu, J., Yang, H., He, H., Wang, W., Tuo, Z., Cheng, W. H., ... & Fu, J. (2023, October). Moviefactory: Automatic movie creation from text using large generative models for language and images. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9313-9319).
[22] Ni, M. (2024, July). Research on the Application of Reinforcement Learning in Film Script Generation and Narrative Innovation. In International Workshop on New Approaches for Multidimensional Signal Processing (pp. 181-193). Singapore: Springer Nature Singapore.
[23] Wang, Y. (2024). Automatic Generation Algorithm of Movie and TV Scripts Based on ChatGPT. International Journal of Maritime Engineering, 1(1), 531-544.
[24] Zheng, H. (2025). Artificial intelligence-driven sentiment analysis and optimization of movie scripts. Discover Artificial Intelligence, 5(1), 114.
[25] Rao, A., Chou, J. P., & Agrawala, M. (2024, October). Scriptviz: A visualization tool to aid scriptwriting based on a large movie database. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (pp. 1-13).
[26] Wei, Z., Wu, H., Zhang, L., Xu, X., Zheng, Y., Hui, P., ... & Rao, A. (2025, September). CineVision: An Interactive Pre-visualization Storyboard System for Director–Cinematographer Collaboration. In Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (pp. 1-18).
[27] Hong, Z., Xu, X., & Liu, X. (2025). Application of virtual reality technology based on artificial intelligence in 3D animated film storyboard. Discover Computing, 28(1), 147.
[28] Hilal, R. (2025, August). AI-Driven Innovation in Film Set Design: Toward Automated and Immersive Virtual Production. In 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) (pp. 1-7). IEEE.
[29] Kavitha, L. (2023). Copyright challenges in the artificial intelligence revolution: Transforming the film industry from script to screen. Trinity Law Review, 4(1), 1-8.
[30] Ju, Y., & Wei, G. (2024). Film and Television Special Effects AI System Integrating Computer Artificial Intelligence and Big Data Technology. Scalable Computing: Practice and Experience, 25(4), 2532-2539.
[31] Tian, Y., & Yang, F. (2025). Deep learning-driven real-time rendering technology for film and television animation special effects. International Journal of Information and Communication Technology, 26(33), 57-75.
[32] Wang, D., & Lathakumari, K. R. (2023, November). Extraction Algorithm of Film and Television Special Effects Based on Artificial Intelligence Technology. In International Conference on Cognitive based Information Processing and Applications (pp. 83-92). Singapore: Springer Nature Singapore.
[33] Zhang, J. (2024). Visual Communication Method of Multi frame Film and Television Special Effects Images Based on Deep Learning. Scalable Computing: Practice and Experience, 25(6), 5460-5468.
[34] Zheng, Y., Cui, J., Zhong, H., & Choi, D. H. (2021, November). Intelligent Repair Method of Old Movie Speckle Noise Based on AI Deep Learning. In 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE) (pp. 53-56). IEEE.
[35] Dong, Q., Zhong, G., & Wu, B. (2024, July). Research on the Digital Restoration of Female Hero Images in Shandong Red Films. In 2024 6th International Conference on Electronics and Communication, Network and Computer Technology (ECNCT) (pp. 524-527). IEEE.
[36] Sun, M. (2025). A method for solving the multiple degradation video quality enhancement problem: a processing framework for AI-based coding damage repair in concert with video super-resolution. Multimedia Systems, 31(1), 52.
[37] Xiao, L. (2025, February). Research and Application of Intelligent Algorithm for Digital Media Video Editing Based on Deep Learning. In 2025 International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC) (pp. 464-469). IEEE.
[38] Jin, C., Lin, M., Wu, F., Wu, X., Zhou, Y., & Wang, J. (2025). TVMTrailer: a text-video-music AIGC framework for film trailer generation. IEEE Transactions on Systems, Man, and Cybernetics: Systems.
[39] Liu, X., & Pan, H. (2022). The path of film and television animation creation using virtual reality technology under the artificial intelligence. Scientific Programming, 2022(1), 1712929.

[40] Shen, W., & Chen, C. (2024, April). 3D Film and Television Animation Production Cloud Management System Based on Artificial Intelligence. In 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 948-953). IEEE.
[41] Rosati, E. (2025). The future of the movie industry in the wake of generative AI: A perspective under EU and UK copyright law. Computer Law & Security Review, 59, 106207.
[42] Alam, M., Khan, S. R., Khan, I. R., & Siddiquee, A. Q. (2024). The Role of Cybersecurity in Digital Film Production and Distribution: Challenges and Solutions. CyberSystem Journal, 1(2), 9-20.
[43] Shang, W., Li, H., Ni, X., Chen, T., & Liu, T. (2025). BlockGuard: Advancing digital copyright integrity with blockchain technique. Computers and Electrical Engineering, 122, 109897.
[44] Wang, F. F. (2022). Resolving Online Content Disputes in the Age of Artificial Intelligence: Legal and Technological Solutions in Comparative Perspective. J. Comp. L., 17, 491.
[45] Qian Zhang,Jia Rui Zhao,Xiao Qian Liu,Yu Wei Zhan,Zhen Duo Chen,Xin Luo & Xin Shun Xu. (2025). Hypergraph-based CLIP hashing for unsupervised cross-modal retrieval. Knowledge-Based Systems,330(PA),114508-114508.
[46] Jiale Cao,Yuanheng Liu,Zhong Ji,Jingren Liu,Aiping Yang & Yanwei Pang. (2025). Mitigating forgetting in the adaptation of CLIP for few-shot classification. Computer Vision and Image Understanding,261,104493-104493.
[47] Lulin Zhang,Ewelina Rupnik,Tri Dung Nguyen,Stéphane Jacquemoud & Yann Klinger. (2025). BRDF-NeRF: Neural radiance fields with optical satellite images and BRDF modelling. International Journal of Applied Earth Observation and Geoinformation,143,104747-104747.