Vol. 35, No. 1, 2026
(this issue contains accepted papers online and is not closed yet)

Machine
GRAPHICS & VISION

International Journal

Published by
The Institute of Information Technology
Warsaw University of Life Sciences – SGGW
Nowoursynowska 159, 02-776 Warsaw, Poland
in cooperation with
The Association for Image Processing, Poland – TPO

Enhancing Cultural Heritage Digitalization through 3D Graphics Algorithm and Immersive Visual Communication Technology

Fang Yuan*

Guangxi Normal University for Nationalities, College of Art, Chongzuo, China

**Corresponding author: Fang Yuan (yuanfang16316@163.com)*

Submitted: 17 Jun 2025 **Accepted:** 06 Oct 2025 **Published:** 16 Feb 2025

Abstract With the continuous advancement of digital technology, cultural and creative product design is shifting from static presentation to dynamic immersive experience. The research aims to address the challenges faced by traditional modeling methods in accurately restoring complex textures and cross platform visual communication. The neural radiation field algorithm was enhanced by introducing a multi-level cost volume fusion module and a Gaussian uniform mixture sampling strategy. Furthermore, a collaborative visual communication framework integrating augmented reality and virtual reality was constructed, achieving a transition from single image input to high-precision 3D reconstruction, and then to dynamic interaction. The experiment showed that the improved algorithm achieved peak signal-to-noise ratios of 30.63 and 30.15 on the UoM-Culture3D and Bootstrap 3D synthetic datasets, respectively, with structural similarity indices of 0.88 and 0.89, respectively. Field deployment tests have shown that integrating AR and VR technologies into visual communication strategies significantly improves spatial perception consistency, prolongs user engagement time, and enhances detail recognition accuracy. This research emphasizes the potential of combining deeply coupled 3D graphics algorithms with immersive technology, which can help improve the digital restoration accuracy and cultural dissemination efficiency of cultural and creative products, thereby supporting the modern inheritance of traditional culture.

Keywords: 3D graphics algorithm, visual communication technology, cultural and creative product design, NeRF, VR, AR.

1. Introduction

With the adoption of digital technology in the industry, cultural and creative product design is facing a transition from static output to dynamic immersive experience. The popularization of Virtual Reality (VR) hardware and the advancement of real-time graphics computing have made the digital revitalization of cultural heritage a new direction. Through technological means, it is possible to break through the physical limitations of physical exhibitions, allowing historical patterns and traditional techniques to gain cross temporal and spatial dissemination power. This cultural and creative product design has put forward new requirements and urgently needs to break through the limitations of traditional two-dimensional expression, establish a multidimensional design system that integrates high-precision modeling, dynamic narrative, and interactive experience [14]. Currently, the Neural Radiation Field (NeRF) technology in the field of 3D graphics algorithms combines ray tracing and deep learning to achieve high fidelity digital reconstruction of complex cultural carriers such as cultural relics patterns and historical scenes [21]. However, this technology relies on dense input of hundreds of images in a single scene and time-consuming training on a scene by scene basis, making it difficult to adapt to the fast iterative design process of cultural and creative products [12, 31]. The augmented reality (AR) and VR technologies in the field of visual communication can create a virtual real fusion experience environment. However, most of the existing schemes use a single mode, which has problems such as large spatial alignment error, homogenization of interaction forms, and shallow semantic analysis of cultural symbols [27, 32]. To this end, a multidimensional design method for cultural and creative products based on the Improved NeRF (INeRF) algorithm and the integration of AR and VR is proposed. By integrating multi-level cost structures and utilizing cross-scale feature fusion techniques, geometric reasoning capabilities are strengthened. Furthermore, the implementation of a Gaussian uniform mixture sampling strategy optimizes the efficiency of surface detail reconstruction. Consequently, a seamless interactive experience across AR and VR platforms is attained within the visual communication layer. The research aims to enhance the cultural connotation expression and user experience of cultural and creative products, and promote the development of the cultural and creative industry towards digitalization and multidimensionality. The innovation of the research lies in introducing a multi-level geometric feature fusion mechanism and a mixed sampling strategy into the NeRF framework. Meanwhile, through AR-VR collaborative interactive design, the organic unity of cultural symbols in spatial, temporal, and perceptual dimensions is achieved, providing practical and expressive methodological support for the digital innovation of cultural and creative products.

2. Related works

High-precision 3D reconstruction is the cornerstone of cultural heritage digitization. NeRF technology has garnered significant attention for its ability to fuse ray tracing with deep learning, enabling high-fidelity reconstruction of the complex textures and structures of cultural relics. However, classical NeRF and its variants generally suffer from significant limitations: their training process heavily relies on hundreds of dense multi-view images from a single scene and time-consuming scene-by-scene optimization, which severely restricts their applicability in cultural and creative product design workflows requiring rapid iteration. To address reconstruction challenges in specific domains, researchers have proposed various optimization schemes. To achieve texture synthesis optimization, Houdard et al. [9] proposed a general framework named GOTEX. By constraining the local feature statistical distribution and utilizing the optimal transport semi-dual formula to control the feature distribution, high-quality texture synthesis and restoration were achieved. To improve the accuracy and efficiency of 3D reconstruction of ancient buildings, Ge et al. [7] introduced depth supervision into the NeRF framework, combining a truncated signed distance function and an incremental training strategy, effectively enhancing the accuracy and efficiency of 3D reconstruction of ancient buildings. In the field of dynamic scene reconstruction, Qiu et al. [19] innovatively combined NeRF with signed distance fields to achieve realistic reconstruction of dynamic ship models, demonstrating its potential for dynamic modeling of specific objects. Mazzacca et al. [15] further validated the effectiveness of NeRF in reconstructing cultural heritage datasets, particularly in handling uniform textures or shiny surfaces, expanding the documentation pathways for digital heritage.

Visual communication technology serves as a bridge connecting digital reconstruction outcomes with user experience. AR and VR technologies enable the creation of immersive cultural experience environments that blend virtual and real elements. To enhance the visual communication effectiveness of digital animated advertisements, Fang et al. [5] proposed a multimodal visual communication system model based on multimodal video emotion analysis. This model dynamically adjusts digital animated advertisement content according to user emotions, enhancing the personalization and appeal of interactions, and demonstrating the potential of emotion-driven content adaptation. Liu et al. [13] conducted an in-depth analysis of visual communication strategies for cultural imagery in rural environments, emphasizing the importance of environmental perception in experiencing cultural spirit through the integration of art intervention institutions, and providing insights for cultural narratives in specific spaces. In terms of communication effectiveness evaluation, the video data analysis system by Yachnaya et al. [26] can identify and assess paralinguistic and non-verbal components in communication, providing tools for quantifying user experience. Yudhanto et al. [30] advocate a visual communication design philosophy grounded in culture and communication, emphasizing the importance of researching the target audience's values, norms, language, beliefs, and visual elements to enhance the cultural relevance and effectiveness of design.

As can be seen from the above, although three-dimensional graphics algorithms and visual communication technologies have made significant progress in their respective fields, there remains a lack of cross-platform, multi-modal integrated design methods for the digitization of cultural heritage. Existing solutions often struggle to balance high-precision texture restoration, real-time interactive performance, and visual consistency across multiple devices. This research gap leads to issues such as experience discontinuity and information loss in the actual dissemination of cultural and creative products. To address this, the study proposes a multi-dimensional design framework for cultural and creative products based on an improved INeRF and the deep integration of AR and VR, providing a solution that combines precision and expressiveness for the innovative transformation and dissemination of cultural heritage.

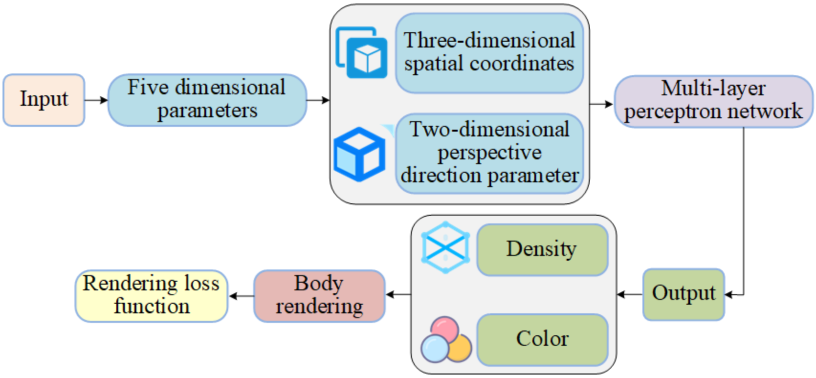

Input
Five dimensional parameters
Three-dimensional spatial coordinates
Two-dimensional perspective direction parameter
Multi-layer perceptron network
Output
Density
Color
Body rendering
Rendering loss function

Fig. 1. Schematic diagram of NeRF algorithm (icons designed by Freepik [6]).

3. Methods and materials

3.1. Design of 3D graphics algorithm based on INeRF

Three-dimensional graphics algorithms are driving the transformation of cultural and creative products toward multi-dimensional design. NeRF technology combines deep learning and ray tracing to achieve high-fidelity three-dimensional reconstruction of cultural relics and historical scenes, effectively restoring complex textures and material effects, and solving the challenges of traditional modeling in reproducing complex materials and intricate patterns [11, 16, 28]. The basic structure of NeRF technology is shown in Fig. 1. This technology first receives a five-dimensional input parameter consisting of spatial position coordinates and the angle of light incidence. This parameter is then mapped by a multi-layer perceptron network into RGB color values and density parameters. Subsequently, the system emits rays from the viewpoint, continuously sampling points along the path, and uses a volume rendering formula to calculate the transmittance and color contribution of each point, thereby synthesizing a realistic lighting effect. Finally, the model is optimized using a pixel-level rendering loss function to approximate the optical properties of the real-world scene. Among them, the NeRF mapping function [10] is

$$F(x, y, z, \theta, \phi) \to (R, G, B, \sigma)\,, \tag{1}$$

where x, y, and z represent three-dimensional spatial coordinates, θ and ϕ represent the angle parameters of the incident direction of light rays, R, G, and B represent the RGB color values of the sampling points, and σ represents the medium density of the sampling point. The function predicts the optical properties of each sampling point based on the light and scene geometry characteristics, thereby providing basic data for

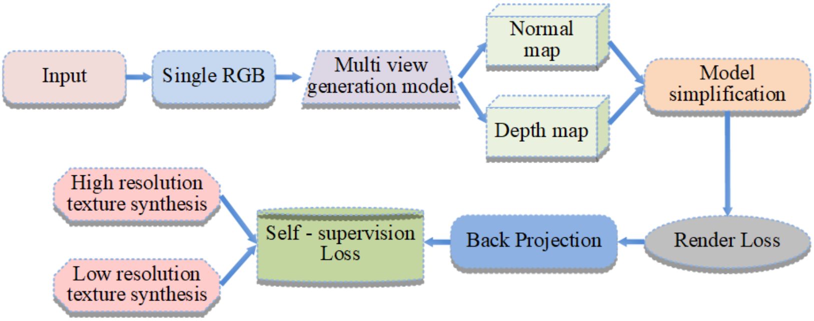


Fig. 2. The basic framework of INeRF.

volume rendering. The rendering expression is

$$C(r) = \int_{t_n}^{t_f} T(t) \cdot \sigma(r(t)) \cdot c(r(t), d)\, \mathrm{d}t\,, \tag{2}$$

where $C(r)$ represents the cumulative color of light, $T(t)$ indicates the transmittance of light from the starting point to the current point, $\sigma(r(t))$ represents the density of path point $r(t)$, $c(r(t), d)$ indicates the color of the path point $r(t)$ in the direction d, and t_n and t_f represent the starting and ending points of the light. This formula achieves optically realistic image synthesis by accumulating the color and transparency of each sampling point along the light path. The expression of the rendering loss function is

$$\mathcal{L}_{\text{render}} = \sum_{p} \left\| \hat{C}(p) - C_{\text{gt}}(p) \right\|^2 , \tag{3}$$

where $\mathcal{L}_{\text{render}}$ represents pixel-level rendering loss, $\hat{C}(p)$ represents the color of pixels in the generated image, and $C_{\text{gt}}(p)$ represents the color of pixel p in real multi-view images. Although NeRF technology can achieve high-precision 3D reconstruction, it relies on a large number of input images from a single scene and time-consuming scene by scene optimization training, which makes it difficult to meet the design requirements for rapid iteration of cultural and creative products. Therefore, the study proposed the INeRF algorithm, whose basic framework is shown in Figure 2.

The INeRF algorithm starts with a single RGB input and extends the model to multi view data through multi view generation. It combines camera parameters to drive the 3D reconstruction module to generate normal maps and depth maps. During the process, supervised and soft supervised loss optimization is used to optimize depth and RGB prediction, and geometric consistency is ensured through backprojection. The rendering loss function further optimizes the lighting and material performance of the

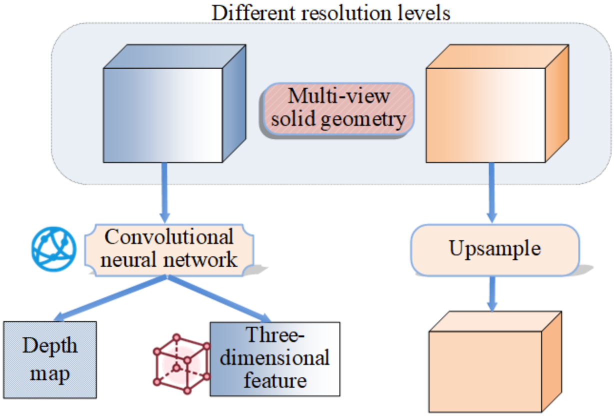

Different resolution levels
Multi-view solid geometry
Convolutional neural network
Upsample
Depth map
Three-dimensional feature

Fig. 3. The basic structure of the cost body (icons designed by Freepik [6]).

model, followed by high-resolution texture synthesis to enhance details, and ultimately balances accuracy and efficiency through model simplification techniques to output high-quality 3D models.

To address the issue of insufficient geometric information in single-view input, a multi-level cost volume fusion module based on convolutional attention was designed, as shown in Figure 3. Its the schematic diagram is based on multi-view solid geometry. Firstly, feature maps are extracted from input images of different resolution levels, and the three-dimensional geometric information of the scene is captured by constructing a multi-scale cost volume. In the feature fusion stage, low-resolution cost bodies encode global semantics, while high-resolution cost bodies retain details. Cross-layer feature interaction is achieved through convolutional attention, and channel and spatial attention are used to optimize the coordination of local and global information. Ultimately, a geometric neural field with both spatial accuracy and semantic integrity is formed, providing multi-level feature support for rendering. The formula for multi-level cost volume fusion is

$$F_{\text{fused}} = \sum_{l=1}^{L} w_l \cdot F_l^{\text{up}} + F_{\text{res}} , \tag{4}$$

where F_{fused} represents the fused multi-level features, F_l^{up} represents the features after upsampling at layer l, w_l represents the feature weight calculated through attention mechanisms, F_{res} represents the residual connection feature, and L indicates the total number of feature levels. In the feature decoding and rendering optimization stage, IN-eRF achieves efficient and accurate volume rendering by improving the sampling strategy and loss function design. In response to the problem of insufficient density in traditional uniform sampling, a Gaussian uniform mixture sampling strategy is proposed. Based on the depth prior information inferred from multi-view solid geometry, Gaussian distribution dense sampling is used in the surface area of the object, while maintaining uniform sampling density in non critical areas. The expression for Gaussian uniform mixture sampling distribution is [18]

$$P(s) = \lambda \cdot \mathcal{N}(s \mid \mu_d, \sigma_d) + (1 - \lambda) \cdot \mathcal{U}(s \mid s_{\min}, s_{\max})\,, \tag{5}$$

where $P(s)$ represents the probability density function of the sampling point s, $\mathcal{N}$ is the Gaussian distribution, $\mathcal{U}$ represents the uniform distribution, λ represents mixed weight coefficients, μ_d represents the depth mean, and σ_d represents variance.

Meanwhile, a deep self-supervised loss function was designed to generate pseudo depth maps using multi-view consistency constraints. The pixel information of the source view was distorted to the target perspective through differentiable reprojection, and a self-supervised signal without the need for real depth annotation was constructed. Moreover, during the feature decoding stage, the algorithm spatially aligns the three-dimensional local features generated by the geometric neural field with the two-dimensional global features. It then incorporates the encoded information of light ray directions, dynamically decoding the color and density values for each sampling point via a multi-layer perceptron. Finally, it synthesizes the pixel color and depth information of the target viewpoint using a differentiable rendering equation, thereby establishing an end-to-end trainable framework. Through this framework, designers can quickly convert historical images, physical photos, or 2D drawings into interactive 3D models, greatly improving the responsiveness and flexibility of the creative production process.

The pseudocode of the INeRF algorithm is presented in the Appendix A.

3.2. Design of cultural and creative products based on visual communication technology

After completing high-precision digital reconstruction based on 3D graphics algorithms, visual communication technology has become the core supporting means in multi-dimensional expression of cultural and creative products. To achieve deep dissemination and innovative expression of cultural values, a deep integration strategy based on AR and VR has been studied and designed. The overall framework is shown in Figure 4. In the data generation layer, the system relies on the INeRF algorithm to construct a high-precision 3D model from a single image, obtaining multidimensional data including geometry, normal maps, and depth maps, laying the foundation for subsequent visual presentation. The visual expression layer focuses on the graphic rendering and semantic visualization processing of 3D models, mapping digital models into recognizable and culturally significant visual content through lighting simulation, material mapping, and color coding, and adapting to AR and VR platforms for dynamic presentation [29].

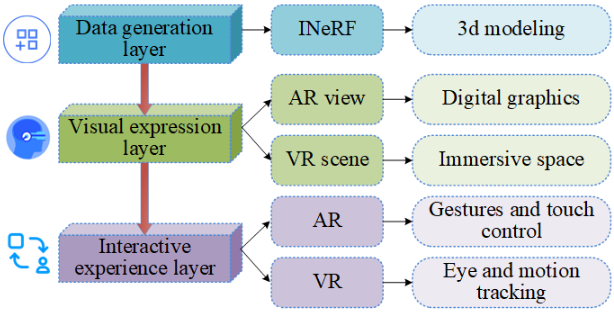


Fig. 4. The overall framework of visual communication strategy of cultural and creative products (icons designed by Freepik [6]).

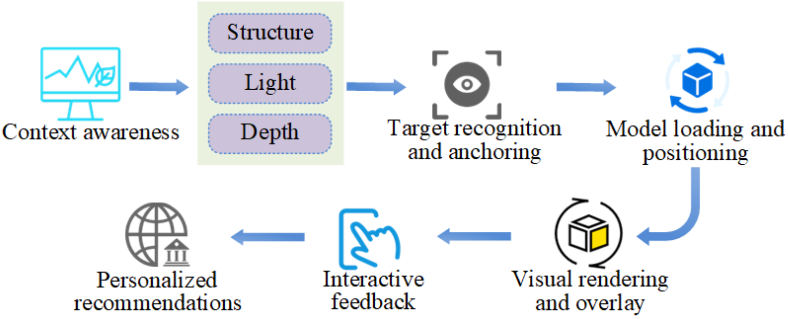

Context awareness
Structure
Light
Depth
Target recognition and anchoring
Model loading and positioning
Visual rendering and overlay
Interactive feedback
Personalized recommendations

Fig. 5. Visual communication flow chart of AR-based cultural and creative products (icons designed by Freepik [6]).

The interactive experience layer revolves around user perception, combining the real-time positioning and virtual real overlay capabilities of AR, as well as the immersive spatial construction characteristics of VR, to achieve dynamic calling and multi-modal interaction design of cultural and creative graphic content.

The basic process of visual communication for AR-based cultural and creative products is shown in Figure 5. The system uses the RGB-D sensor built into the AR device to collect data on the geometric structure, depth distribution, and lighting conditions of the user's surroundings. It then uses feature point matching algorithms to identify and anchor targets, accurately locating physical objects such as display cases, cultural and creative packaging, and interior walls, and setting attachment points for virtual elements [22]. During the graphic deployment phase, the 3D models generated by INeRF are compressed and optimized for lightweight performance, then loaded into the augmented reality platform. The system automatically adjusts the orientation based on the on-site

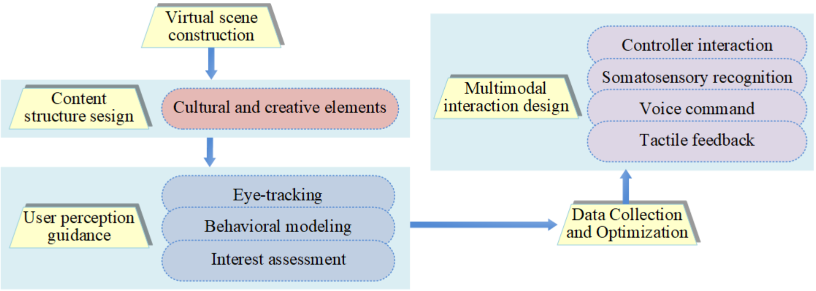


Fig. 6. Framework diagram of cultural and creative space construction based on VR (icons designed by Freepik [6]).

coordinate system. Subsequently, the system performs real-time graphic rendering and visual overlay, utilizing dynamic lighting estimation and reflection maps to ensure high consistency between virtual images and the real-world environment. Users can interact with virtual graphics through gesture recognition, voice input, or touch operations to obtain multi-dimensional feedback. The system finally combines user behavior trajectories and preference patterns to achieve personalized push notifications for cultural and creative content, further enhancing the targeting and engagement of visual communication [4]. To achieve seamless integration between virtual and real environments, the study developed an AR-VR hybrid interaction framework. When users transition from an AR scene to a VR scene, the system retains their operational state and interaction history through a spatial state caching mechanism, enabling state restoration and content continuity within the virtual space. First, the system uses the RGB-D sensor and IMU data from AR devices for real-time environmental mapping and user localization. Second, a virtual scene mapping model is established on the VR platform to ensure that the scene geometry aligns with the real-world spatial coordinates [3]. Finally, a state caching and synchronization mechanism is designed to save user interaction operations and object states, enabling seamless cross-device switching. In terms of on-site deployment, the system considers lighting matching, dynamic occlusion handling, and device load optimization to ensure stable operation in exhibition or cultural and creative experience spaces.

The AR-VR hybrid interaction framework is shown in Figure 6, which is the framework for constructing cultural and creative spaces based on VR. It systematically outlines the methodological path of VR technology in multidimensional cultural and creative design. Firstly, the designer relies on a 3D model database and INeRF generated results to construct a virtual environment that covers historical block restoration, cultural festival scenes, and immersive exhibition spaces for cultural relics, forming a virtual field with cultural depth. At the level of content structure, cultural and creative elements are orderly embedded into spatial nodes, forming multiple types of information units, including decorative shapes, interactive objects, semantic labels, and dynamic animations, thus establishing a rich cultural narrative space. The system integrates gaze tracking and behavior modeling modules to dynamically adjust the visual hierarchy and dynamic parameters of virtual content based on users' attention paths and interest preferences, guiding users to naturally integrate into the narrative process. In terms of interaction, the platform integrates controller control, speech recognition, motion capture, and tactile feedback technology to provide users with multi-channel immersive interaction methods, enhancing the degree of freedom and realism of the experience. Meanwhile, the system continuously collects user behavior data in the virtual space in the background, including field of view movement, dwell time, and interaction frequency, providing data support and model basis for subsequent scene structure adjustment and visual information optimization, thereby achieving iterative updates and precise push of the design system.

4. Results

4.1. Performance verification of 3D graphics algorithm based on INeRF

To verify the effectiveness of multi-dimensional design of cultural and creative products based on 3D graphics algorithms and visual communication technology, a 3D reconstruction and visual communication system for cultural and creative products based on INeRF algorithm and AR/VR fusion was constructed in an experimental environment with GPU acceleration capability.

The image datasets used in the experiment included the UoM-Culture3D dataset [25] and the Bootstrap3D synthetic dataset [23, 24]. The UoM-Culture3D dataset contains multi-perspective images of historical artifacts and cultural scenes, with a resolution of 1920 × 1080, suitable for high-quality 3D reconstruction. The Bootstrap3D synthetic dataset contains millions of multi-view images covering creative objects such as fictional creatures and cultural symbols.

The specific experimental environment and parameter configuration are shown in Table 1. Based on this experimental environment, the study compared the introduction of raw NeRF [16,28], NeRF based on multi-resolution texture pyramid (Mip-based, Mip-NeRF) [2], Instant Neural Graphics Primitives with a multi-resolution hash encoding (Instant-NGP) [17], and INeRF model proposed in this paper.

Firstly, using Peak Signal to Noise Ratio (PSNR) as a comparison metric, tests were conducted on different datasets, and the results are shown in Figure 7, where the PSNR comparison performance of four 3D reconstruction models on two datasets are displayed. In Figure 7a, on the UoMCult3D dataset, NeRF had the weakest performance with a PSNR of 25.82 at the 500th iteration. Mip-NeRF and Instant-NGP reached 28.19 and

Tab. 1. Experimental environment and parameter configuration.

Type	Name	Version
Hardware equipment	CPU	Intel Xeon Gold 6248R, 3.0 GHz, 24C
	GPU	NVIDIA RTX 3090, 24 GB RAM
	RAM	128 GB DDR4
	Memory device	2 TB NVMe SSD
Software equipment	Operating system	Ubuntu 20.04 LTS
	DL framework	PyTorch 1.13
	Graphics rendering	Unity 2022.3 (HDRP line pipe)
	AR develop	ARCore 1.35, ARKit 5.0
	VR develop	SteamVR 2.0, OpenXR 1.0
Parameter name	Learning rate	0.001
	Batch size	1024
	Render resolution	800 × 800 pixels
	Real-time render target frame rate	$\geq$ 30 FPS

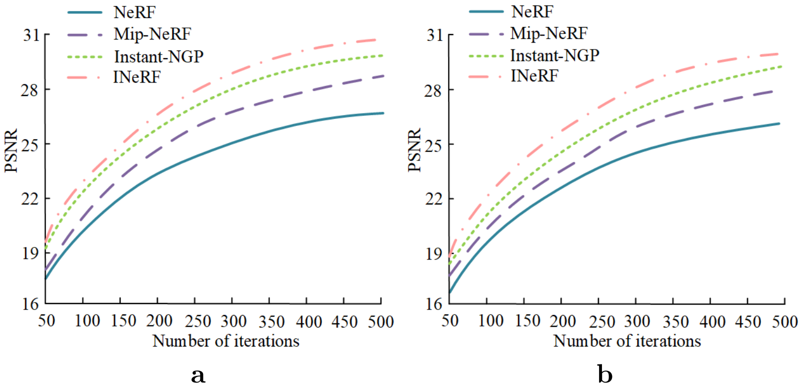


Fig. 7. PSNR comparison of four models with different data sets: (**a**) UoM-Culture3D, (**b**) Bootstrap3D.

30.11, respectively, while INeRF performed the best, stabilizing at 30.63, with an average improvement of 9.24% compared to the other three models. In Figure 7b, INeRF still had a significant advantage on the Bootstrap3D dataset, with a PSNR of 30.15 at the 500th iteration, an average increase of 8.17% compared to other models. This indicated that INeRF had good universality and reconstruction stability in stylized data and cultural images. On this basis, the graphic loading speed and Root Mean Square Error (RMSE) of four models on the AR platform were tested, and the results are shown in Figure 8. According to Figure 8a, as the number of experiments increased, the loading speed of

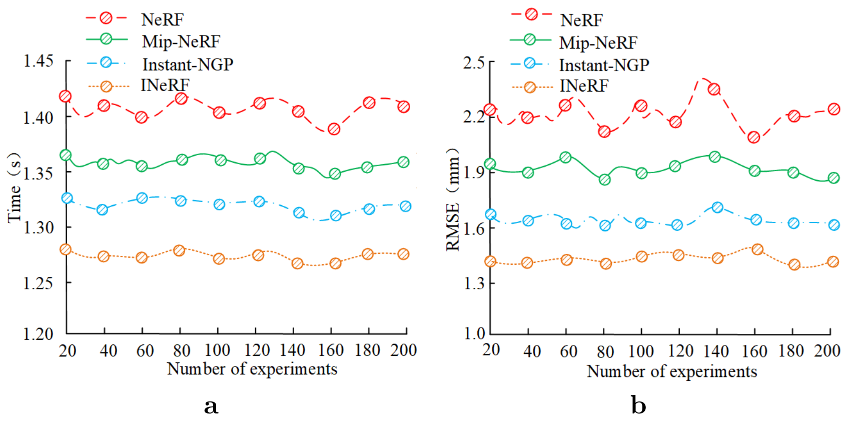


Fig. 8. Comparison of parameters of four models: (**a**) graphic loading time, (**b**) RMSE.

the INeRF model remained at a relatively low level of about 1.26 s, demonstrating high stability and efficiency. In contrast, the NeRF model had the longest loading time, close to 1.41 seconds, and it fluctuated greatly. This might have been due to its reliance on a large number of input images and scene-by-scene optimization training, which resulted in high computational complexity and a slow speed during the loading process. Based on Figure 8b, INeRF had the lowest RMSE value among 200 experiments, stabilizing at around 1.42 mm, with an average reduction of 25.28% compared to other models. Overall, the balance between speed and accuracy of INeRF validated the effectiveness of its improved architecture, providing a reliable technical path for high-fidelity digitization of cultural heritage.

Meanwhile, the Structural Similarity Index Measure (SSIM) of four models on different datasets were compared, and the results are shown in Figure 9. Figure 9a shows the SSIM comparison of four models on the UoM-Culture3D dataset. As the number of iterations increased, the SSIM value of INeRF gradually rose and tended to stabilize. When the number of iterations reached 500, the SSIM value of INeRF remained stable at around 0.88, significantly better than the other three models. Figure 9b presents the SSIM comparison of four models for the Bootstrap3D dataset. NeRF performed better than other models on the Bootstrap3D dataset. When the number of iterations reached 500, the SSIM value of INeRF reached 0.89. This indicates that INeRF can effectively integrate geometric features of different scales, enhancing the model's perception and reconstruction ability of complex image structures.

To directly validate the accuracy of the INeRF algorithm in 3D structure reconstruction, a quantitative evaluation based on point cloud comparison was conducted on the

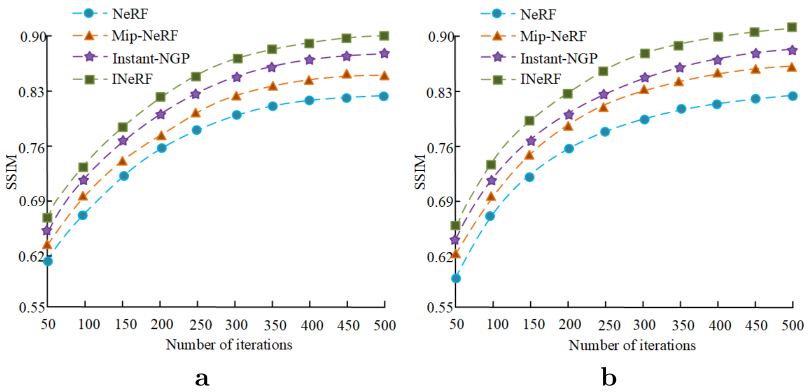


Fig. 9. Comparison of SSIM of four models for different data sets: (**a**) UoM-Culture3D, (**b**) Bootstrap3D.

Tab. 2. Comparison of 3D geometric reconstruction effects of different 3D reconstruction techniques. Asterisks '*' and '**' indicate statistically significant differences compared to INeRF at $p < 0.05$ and $p < 0.01$, respectively.

Model	NeRF	Mip-NeRF	Instant-NGP	INeRF	3DGS	DiffRF
Chamfer dist. [mm]	2.12*	1.88*	1.55*	1.42	1.60	1.50
Hausdorff dist. [mm]	6.48**	5.92**	5.11*	4.78	5.05	4.92
F1-score @0.05	0.42**	0.51**	0.61*	0.65	0.62	0.63
F1-score @0.1	0.59**	0.65**	0.74*	0.78	0.75	0.76
F1-score @0.2	0.71**	0.76**	0.83*	0.86	0.84	0.85
Normal Consistency	0.78**	0.81**	0.85*	0.88	0.86	0.87
Training time [h]	12.42	9.71	4.15	5.31	6.27	6.82
Peak vRAM [GB]	18.60	16.24	9.83	11.41	12.58	13.16

UoM-Culture3D dataset. Two emerging 3D reconstruction techniques were also introduced for comparison: the 3D Gaussian Splatting (3DGS) model and the Rendering-Guided 3D Radiance Field Diffusion Model (DiffRF). Marching Cubes algorithm was used to extract meshes from the density fields predicted by each model, and 50 000 vertices were uniformly sampled to generate point clouds for evaluation. The results are shown in Table 2. It can be seen that the NeRF model performs the worst in various indicators, reflecting its insufficient ability to reconstruct complex textures and details in sparse views, as well as high resource requirements. Mip NeRF improved feature expression through multi-resolution texture pyramids, reducing Chamfer Distance to

1.88 mm and Hausdorff Distance to 5.92 mm. However, there were still significant differences ($p < 0.05$) between the improvements and INeRF. Instant NGP further optimized the point cloud distribution under dense feature encoding, with a Chamfer Distance of 1.55 mm and a normal consistency of 0.85. Although the overall accuracy is close, the difference with INeRF is still significant ($p < 0.05$). In contrast, INeRF achieved the best performance on all indicators, with the lowest Chamfer Distance being 1.42 mm, the Hausdorff Distance dropping to 4.78 mm, F1-scores reaching 0.78 and 0.86 at the 0.1 and 0.2 thresholds, respectively, and a normal consistency of 0.88. The INeRF maintains high accuracy while controlling the training time to 5.31 h, with a video memory usage of only 11.41 GB. Although slightly higher than Instant NGP, it still demonstrates good deployability in resource constrained environments, reflecting the balance advantage between accuracy and efficiency. The difference from most methods is significant or highly significant, thanks to the collaborative optimization of multi-level cost volume fusion and Gaussian uniform mixture sampling strategy in details and global structure. For emerging technologies, the 3D Gaussian jet model and rendering guided radiation field diffusion model approach Instant NGP on Chamfer Distance and F1-score, with no significant difference compared to INeRF, but slightly lower in performance, indicating that there are still subtle geometric errors in sparse input and complex texture scenes.

Based on various indicators and statistical analysis, INeRF exhibits excellent performance in point cloud accuracy, surface normal consistency, and F1-score at different scales. It also shows strong advantages in computational resource utilization, verifying its robustness and reliability in high-precision 3D reconstruction. At the same time, it demonstrates strong adaptability to complex textures and geometric structures in the process of cultural heritage digitization.

4.2. Visual communication effect verification

To validate the effectiveness of the proposed visual communication strategy integrating AR and VR in actual deployment, the study conducted on-site deployment tests in museum exhibition spaces. The deployment included: AR end: Using ARCore/ARKit devices to scan the exhibition area, accurately anchor the location of exhibits, and overlay virtual information. VR end: Using SteamVR devices to construct virtual exhibitions of historical scenes, allowing users to freely interact in the virtual space. The actual measurement data covers indicators such as spatial perception consistency, interaction fluidity, immersion, and cultural understanding perception (out of 10 points) for 30 test subjects. The study compared the traditional 2D display, single VR, and single AR technologies with the proposed fusion strategy, and tested the spatial perception consistency and average dwell time of the four technologies in four cultural and creative scenes: porcelain, murals, ancient architecture, and bronze ware. The results are shown in Figure 10.

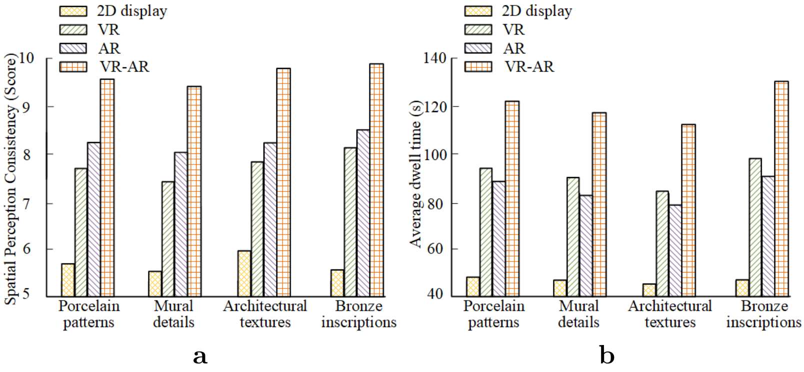


Fig. 10. Comparison of visual communication effects in different cultural and creative scenes: (**a**) spatial consistency ratings; (**b**) average dwell time.

According to Fig. 10a, the scores for integrating VR–AR technology in the four cultural and creative scenes of porcelain, mural, ancient architecture, and bronze ware were 9.50 points, 9.40 points, 9.60 points, and 9.83 points, respectively. When compared with the other three single visual communication technologies, the average scores had increased by 32.72%, 38.35%, 32.27%, and 34.25%, respectively. This indicated that the integration of VR–AR technology achieved better real-world mapping and spatial positioning of three-dimensional structures under the fusion of virtual and real environments. Meanwhile, based on Fig. 10b, the average dwell time of the fusion strategy in the four cultural and creative scenes of porcelain, mural, ancient architecture, and bronze ware was 121.24 s, 118.16 s, 115.67 s, and 130.13 s, respectively. This was an average increase of 59.67%, 58.32%, 57.19%, and 62.25% compared to the other three technologies. By constructing an immersive virtual space and implementing a personalized interactive content push mechanism, users were able to form a deeper sense of participation and cultural context immersion during the experience, which in turn extended their stay time.

Finally, the interactive experience and cultural perception effects of visual communication technology integrating AR and VR were studied, and the results are shown in Table 3. Visual communication technology that integrates VR and AR significantly outperforms single 2D display, VR, or AR solutions in terms of scene detail recognition, interaction fluidity, visual immersion, cultural compatibility, and memory retention. Most of these differences are highly significant ($p < 0.01$), confirming its advantages. Specifically, the scene detail recognition accuracy of the proposed fusion VR and AR visual communication technology reached 92.36%, an average improvement of 23.37% compared to the other three methods, indicating that it had higher accuracy in visual clarity and spatial recognition. In terms of interaction fluency and visual immersion, Tab. 3. Comparison of interactive experience and cultural perception effect of different visual communication technologies. Asterisks ‘*’ and ‘**’ indicate statistically significant differences compared to INeRF at $p < 0.05$ and $p < 0.01$, respectively.

Index	Scene detail recognition accuracy [%]	Interaction fluency [points]	Visual immersion [part]	Cultural fit [%]	Memory retention [%]
2D display	64.37**	5.18**	4.92**	61.25**	58.63**
VR	78.45**	7.86*	7.32**	76.12**	71.40**
AR	81.78*	7.12**	8.47*	80.56*	74.93*
VR–AR	92.36	9.14	9.68	90.42	86.71

the fusion strategy achieved scores of 9.14 and 9.68, respectively, with an average improvement of 36.07% and 39.94% compared to the other three methods. This indicated that it had advantages in operational response and system feedback, while also providing a more immersive cultural experience. In addition, the cultural fit and memory retention of fusion technology were 90.42% and 86.71%, respectively, with an average improvement of 24.47% and 26.92%, indicating that it was more accurate in conveying cultural connotations and symbol fit, and had a stronger effect on retaining cultural information. Overall, the integration of VR and AR technology had significant advantages in enhancing user immersion, improving cultural understanding and memory retention, which validated the scientific and practical nature of the visual communication strategies proposed in the study.

5. Discussion

The research is dedicated to addressing the challenge of synergistically optimizing high-precision reconstruction and cross-platform immersive communication in the digitization of cultural heritage. While 3D reconstruction technology has made progress in multiple fields, it still faces limitations in scene adaptability: a new real-time 3D reconstruction framework significantly enhances maritime situational awareness by integrating temporal 2D video data. Its optimized dynamic reconstruction pipeline enables real-time computation on GPU-accelerated embedded devices. However, it lacks the ability to predict the pose of semi-static objects, making it difficult to capture the geometric continuity of cultural relics under micro-movement conditions [20]. Visual tracking technology based on real-time localization and mapping serves as the core support for augmented reality localization. While it can real-time obtain user pose information, it faces inherent limitations in static scenes due to global localization drift and translation dependency, leading to insufficient spatial anchoring stability in cultural heritage sites [1]. In the field of medical imaging, three-dimensional reconstruction methods for brain tumors based on magnetic resonance imaging demonstrate efficient and precise visualization capabilities. However, when faced with the multi-layered composite texture structure of cultural relics, their topological adaptability remains weak [8]. The aforementioned technologies are either constrained by the integrity of dynamic modeling, limited by the robustness of static localization, or lack the generalization capability for heterogeneous structures, and thus fail to bridge the dual demands of millimeter-level precision reconstruction and multi-modal immersive narrative in cultural heritage digitization.

Therefore, this study aims to establish an integrated system that combines high-precision digital reconstruction with immersive cultural communication, proposing the INeRF algorithm and a multi-dimensional design method that integrates AR and VR technologies. By introducing a multi-level cost-volume fusion module, it achieves collaborative optimization of geometric features across scales, and adopts a Gaussian-uniform hybrid sampling strategy to enhance computational efficiency. Additionally, it combines AR and VR technologies to construct a three-tier communication system encompassing data generation, visual expression, and interactive experience. At the technical implementation level, the system uses multi-sensor fusion to achieve real-time positioning and environmental perception. It also uses dynamic lighting matching, object posture adjustment, and content stream optimization to ensure the accurate presentation of virtual objects on different platforms and in different exhibition environments. Finally, the study validated the feasibility of the integrated AR–VR strategy through field deployment. Field tests demonstrated that the system could achieve stable virtual overlay and multimodal interaction in real exhibition spaces, and user feedback showed significant improvements in cultural information understanding and immersive experiences.

It should be noted that there are still certain limitations in the experimental and validation of the research. Firstly, the test object mainly focuses on the 3D reconstruction of static scenes. However, with the continuous expansion of digital demand for cultural heritage, dynamic cultural heritage such as dance, ceremony, and performance have gradually become research hotspots. For scenes with temporal variability, relying solely on static modeling cannot fully capture their temporal features and dynamic details. Secondly, there are certain limitations to the user research conducted. The current experiment only involves 30 participants, with a relatively limited sample size and a relatively small group composition, which may affect the universality of the research conclusions to some extent and not fully reflect the real experiences of users with different backgrounds.

Future research will further expand the applicability of the INeRF framework in dynamic modeling, such as by introducing temporal consistency constraints and combining optical flow or skeleton driven motion modeling methods to achieve high fidelity reconstruction and presentation of dynamic cultural heritage. At the same time, it is necessary to expand the sample size in user research, increase the dual participation of experts in cultural heritage protection and ordinary visitors, in order to obtain a more comprehensive evaluation. With further validation of the system in multi-user collaboration and dynamic exhibition scenarios, its universality and sustainability in digital protection and cross platform dissemination of cultural heritage are expected to be greatly improved.

6. Conclusion

Compared with existing methods, the INeRF based method improves reconstruction accuracy by 9%, reduces RMSE to 1.42 mm, and enhances visual immersion by nearly 40%. AR–VR integration significantly enhances cultural detail recognition and user engagement. Although research still has limitations in terms of static scene adaptability and small user sample size, future work will explore lightweight network architectures and broader user testing to achieve more universal applications and higher dynamic scene adaptability.

Funding

This paper was supported by the 2024 Guangxi University Research Foundation Capacity Enhancement Project for Middle-aged and Young Teachers: *Research on the Collection, Organization and Digital Inheritance of Endangered Materials on the Feitao Art of Maonan Ethnic Group in Guangxi* (No. 2024KY0765).

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability

The data supporting the findings of this study are referenced in the literature.

References

[1] L. Baker, J. Ventura, T. Langlotz, S. Gul, S. Mills, et al. Localization and tracking of stationary users for augmented reality. *The Visual Computer* 40(1):227–244, 2024. doi:10.1007/s00371-023-02777-2.

[2] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, et al. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5835–5844, 2021. doi:10.1109/ICCV48922.2021.00580.

[3] J. Bast. Managing the image. The visual communication strategy of European right-wing populist politicians on Instagram. *Journal of Political Marketing* 23(1):1–25, 2024. doi:10.1080/15377857.2021.1892901.

[4] J.-J. Cao, S.-M. Fang, and H. Contreras. Multimodal fusion visual communication method based on genetic algorithm. *Journal of Network Intelligence* 10(2):1071–1083, 2025. https://bit.kuas.edu.tw/~jni/2025/vol10/s2/34.JNI-S-2024-05-019.pdf.

[5] J. Fang and X. Gong. Application of visual communication in digital animation advertising design using convolutional neural networks and big data. *Peerj Computer Science* 9:e1383, 2023. doi:10.7717/peerj-cs.1383.

[6] FREEPIK. Find icons that go together. Fast. https://www.freepik.com/icons.

[7] Y. Ge, B. Guo, P. Zha, S. Jiang, Z. Jiang, et al. 3D reconstruction of ancient buildings using UAV images and neural radiation field with depth supervision. *Remote Sensing* 16(3):473, 2024. doi:10.3390/rs16030473.

[8] M. A. Guerroudji, K. Amara, M. Lichouri, N. Zenati, and M. Masmoudi. A 3D visualization-based augmented reality application for brain tumor segmentation. *Computer Animation and Virtual Worlds* 35(1):e2223, JAN 2024. doi:10.1002/cav.2223.

[9] A. Houdard, A. Leclaire, N. Papadakis, and J. Rabin. A generative model for texture synthesis based on optimal transport between feature distributions. *Journal of Mathematical Imaging and Vision* 65(1):4–28, 2023. doi:10.1007/s10851-022-01108-9.

[10] Z. Jia, B. Wang, and C. Chen. Drone-nerf: Efficient nerf based 3D scene reconstruction for large-scale drone survey. *Image and Vision Computing* 143:104920, 2024. doi:10.1016/j.imavis.2024.104920.

[11] X. Liao, X. Wei, M. Zhou, and S. Kwong. Full-reference image quality assessment: Addressing content misalignment issue by comparing order statistics of deep features. *IEEE Transactions on Broadcasting* 70(1):305–315, 2023. doi:10.1109/TBC.2023.3294835.

[12] J. Lin, G. Sharma, and T. N. Pappas. Toward universal texture synthesis by combining texton broadcasting with noise injection in StyleGAN-2. *e-Prime – Advances in Electrical Engineering, Electronics and Energy* 3:100092, 2023. doi:10.1016/j.prime.2022.100092.

[13] F. Liu, B. Lin, and K. Meng. Design and realization of rural environment art construction of cultural image and visual communication. *International Journal of Environmental Research and Public Health* 20(5):4001, 2023. doi:10.3390/ijerph20054001.

[14] W. Liu, Y. Zang, Z. Xiong, X. Bian, C. Wen, et al. 3D building model generation from MLS point cloud and 3D mesh using multi-source data fusion. *International Journal of Applied Earth Observation and Geoinformation* 116:103171, 2023. doi:10.1016/j.jag.2022.103171.

[15] G. Mazzacca, A. Karami, S. Rigon, E. Farella, P. Trybala, et al. Nerf for heritage 3D reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 48(M-2-2023):1051–1058, 2023. doi:10.5194/isprs-archives-XLVIII-M-2-2023-1051-2023.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, et al. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65(1):99–106, 2021. doi:10.1145/3503250.

[17] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics* 41(4):102, 2022. doi:10.1145/3528223.3530127.

[18] M. Pepe, V. S. Alfio, and D. Costantino. Assessment of 3D model for photogrammetric purposes using AI tools based on NeRF algorithm. *Heritage* 6(8):5719–5731, 2023. doi:10.3390/heritage6080301.

[19] S. Qiu, S. Wang, X. Chen, F. Qian, and Y. Xiao. Ship shape reconstruction for three-dimensional situational awareness of smart ships based on neural radiation field. *Engineering Applications of Artificial Intelligence* 136:108858, 2024. doi:10.1016/j.engappai.2024.108858.

[20] F. Sattler, B. Carrillo-Perez, S. Barnes, K. Stebner, M. Stephan, et al. Embedded 3D reconstruction of dynamic objects in real time for maritime situational awareness pictures. *The Visual Computer* 40(2):571–584, 2024. doi:10.1007/s00371-023-02802-4.

[21] S. Shen, S. Xing, X. Sang, B. Yan, and Y. Chen. Virtual stereo content rendering technology review for light-field display. *Displays* 76:102320, 2023. doi:10.1016/j.displa.2022.102320.

[22] X. Shi and R. Villegas. AI technology in the virtual reality environment of graphic design of dynamic art visual communication frame. *Journal of Computational Methods in Sciences and Engineering* 25(3):2603–2616, 2025. doi:10.1177/14727978251321333.

[23] Z. Sun. BS-Objaverse. Hugging Face. `https://huggingface.co/datasets/Zery/BS-Objaverse/`.

[24] Z. Sun, T. Wu, P. Zhang, Y. Zang, X. Dong, et al. Bootstrap3D: Improving multi-view diffusion model with synthetic data. arXiv, arXiv:2406.00093v2, 2024. doi:10.48550/arXiv.2406.00093.

[25] Xinyi_Zheng. CULTURE3D: Cultural Landmarks and Terrain Dataset for 3D Applications. GitHub. `https://github.com/X-Intelligence-Labs/CULTURE3D`.

[26] V. O. Yachnaya, V. R. Lutsiv, and R. O. Malashin. Modern automatic recognition technologies for visual communication tools. *Computer Optics* 47(2):287–305, 2023. doi:10.18287/2412-6179-CO-1154.

[27] C. Yan, B. Gong, Y. Wei, and Y. Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(4):1445–1451, 2020. doi:10.1109/TPAMI.2020.2975798.

[28] J.-W. Yang, J.-M. Sun, Y.-L. Yang, J. Yang, Y. Shan, et al. DMiT: Deformable Mipmapped Triplane representation for dynamic scenes. In: *Computer Vision – ECCV 2024*, pp. 436–453. Springer Nature Switzerland, Cham, 2025. doi:10.1007/978-3-031-73001-6_25.

[29] J. You and X. Lu. Visual communication design based on machine vision and digital media communication technology. *KSII Transactions on Internet & Information Systems* 19(6):1888–1907, 2025. doi:10.3837/tiis.2025.06.007.

[30] S. H. Yudhanto, F. Risdianto, and A. T. Artanto. Cultural and communication approaches in the design of visual communication design works. *Journal of Linguistics, Culture and Communication* 1(1):79–90, 2023. doi:10.61320/jolcc.v1i1.79-90.

[31] Z. Zhang, L. Li, G. Cong, H. Yin, Y. Gao, et al. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7523–7532, 2024. doi:10.1145/3664647.3680777.

[32] M. Zhao. Application of image reconstruction algorithm combining FCN and Pix2Pix in visual communication design. *Journal of Computational Methods in Sciences and Engineering* 25(4):3137–3151, 2025. doi:10.1177/14727978251319398.

A. Appendix

Pseudocode of INeRF algorithm

```
INeRF algorithm
Input:
  $I_{\text{src}}$
  $K$: Camera intrinsic matrix
Output:
  $M$: High-precision 3D mesh model (geometry + texture)
// Step 1: Multi-view generation (replaces multi-image input)
$I_{\text{views}} \leftarrow \text{MultiViewGenerator}(I_{\text{src}})$ //Generate $N$ virtual views $\{I_1, I_2, ..., I_N\}$
$\Theta_{\text{cam}} \leftarrow \text{EstimateCameraPoses}(I_{\text{views}}, K)$ //Estimate virtual view poses
//Step 2: Geometric reasoning (multi-level cost volume fusion)
$F_{\text{multi}} = [\,]$ //Initialize multi-scale feature list
for each $I_i$ in $I_{\text{views}}$
  for each scale $s$ in $[1, 2, 4]$//Multi-resolution feature extraction
    $F_s \leftarrow \text{CNN\_Encoder}(I_i, \text{scale} = s)$ //Extract features at scale $s$
    $F_{\text{multi}}[s] \leftarrow F_s$
  end for
  $C_i \leftarrow \text{BuildCostVolume}(F_{\text{multi}}, \Theta_{\text{cam}}[i])$ //Construct cost volume for view $i$
end for
$F_{\text{fused}} \leftarrow \text{MultiLevelFusion}(C_{\text{all}})$ //Fuse cost volumes (Eq. (4), Fig. 3)
//Step 3: Neural radiance field modeling
for each pixel $p$ in target view:
  ray $r \leftarrow \text{GenerateRay}(p, \Theta_{\text{cam\_target}})$
  //Gaussian-uniform hybrid sampling (Eq. (5))
  samples $\leftarrow$ $\text{GaussUniformHybridSampling}(r, \text{depth\_prior=DepthMap}(F_{\text{fused}}),$
    $\mu = \text{depth\_mean}, \sigma = 0.2, \alpha = 0.7)$ //$\alpha$: Gaussian sampling weight
  $\sigma, c \leftarrow [\,]$ //Store density and color
  for each sample point $x$ in samples
    $\text{feat}_{3d} \leftarrow \text{Query3DFeature}(x, F_{\text{fused}})$ //Query 3D local feature
    $\text{feat}_{\text{dir}} \leftarrow \text{Encode}(\text{view\_dir})$ //View direction encoding
    $(\sigma_x, c_x) \leftarrow \text{MLP}_{\sigma c}(\text{feat}_{3d}, \text{feat}_{\text{dir}})$ //Predict density and color
    $\sigma.\text{append}(\sigma_x)$; $c.\text{append}(c_x)$
  end for
  //Volume rendering (Eq. (2))
  $\hat{C}_p \leftarrow \text{VolumeRendering}(\sigma, c, \text{samples})$
  $\hat{D}_p \leftarrow \text{DepthMapRendering}(\sigma, \text{samples})$ //Predict depth map
end for
//Step 4: Self-supervised optimization
$L_{\text{rgb}} \leftarrow \text{MSE}(\hat{C}, I_{\text{gt}})$ //RGB rendering loss (Eq. (3))
$L_{\text{depth}} \leftarrow \text{DepthConsistencyLoss}(\hat{D}, \text{FusedDepth})$ //Depth self-supervised loss
$L_{\text{total}} \leftarrow \lambda_1 L_{\text{rgb}} + \lambda_2 L_{\text{depth}}$ //$\lambda_1 = 1.0$, $\lambda_2 = 0.5$ (tunable)
Update $\text{MLP}_{\sigma c}$ via $\nabla L_{\text{total}}$ //Backpropagation update
//Step 5: High-res texture generation & model simplification
$M_{\text{highres}} \leftarrow \text{TextureSynthesis}(F_{\text{fused}}, \text{MLP}_{\sigma c})$ //Generate textured dense mesh
$M \leftarrow \text{MeshSimplification}(M_{\text{highres}}, \text{target\_faces=50k})$ //Simplify model
return $M$
```

Intelligent Extraction and Layout Optimization of Digital Media Visual Elements Based on Computer Vision

Hebin Wu*

Department of Computer Engineering, Shanxi Engineering Vocational College, Taiyuan, China

**Corresponding author: Hebin Wu (WuHebin1989@163.com)*

Submitted: 04 Jun 2025 **Accepted:** 09 Dec, 2025 **Published:** 21 Feb 2026

Abstract In the field of digital media, intelligent extraction and layout optimization of visual elements face challenges such as inaccurate semantic understanding of elements and low efficiency in generating layout strategies. This study proposes an extraction and layout optimization model that integrates visual semantic understanding with intelligent optimization strategies, based on a segmentation Vision Transformer and Multi-Objective Firefly Algorithm. The model also utilizes the improved optical flow methods to efficiently capture dynamic information during the design process. Experimental results show that the segmentation Vision Transformer algorithm achieves an extraction accuracy of $98.8 \pm 0.2\%$ for different categories of visual elements. As the training progresses to 50 iterations, the average Intersection-Over-Union stabilizes at 0.95, and the harmonic mean of recall reaches $98.17 \pm 0.38\%$. The evaluation of the integrated model shows that it achieves 99% accuracy in extracting visually similar elements. After layout optimization using the model, the aesthetic score increases to 95.6, and the spatial occupancy rate improves to 97.2%. The above results indicate that the model proposed by the research institute can effectively enhance the accuracy of visual element extraction and the quality of layout optimization, significantly reducing the reliance of traditional methods on manual rules, and providing an efficient and adaptive solution for the automated design of digital media.

Keywords: digital media, layout optimization, SAM, ViT, PWCNet, MOFA.

1. Introduction

In recent years, with the rapid development of the digital media industry, users have raised higher demands for the accuracy and efficiency of visual element processing. The application of computer vision algorithms in the field of digital media not only enables precise extraction of visual elements but also enhances the visual appeal of content through layout optimization [16]. Therefore, research on intelligent extraction and layout optimization algorithms for visual elements is of great significance for technological innovation in the digital media industry. Currently, mainstream visual processing methods have limitations, including poor adaptability to complex scenes, weak dynamic element processing capabilities, and a lack of multi-objective coordination in layout optimization [28]. Traditional methods are prone to incomplete extraction when dealing with dynamic elements in videos, and the layout is also difficult to balance aesthetics and functionality. Specifically, the core research issues that urgently need to be addressed in the current field can be summarized into three points. The first is the disconnection between *segmentation and semantics* in the extraction of static visual elements. Although existing segmentation algorithms can accurately locate the boundaries of elements, they are difficult to capture the semantic associations between elements and cannot directly support layout optimization. Second, the temporal correlation modeling of dynamic visual elements is insufficient. Traditional methods are difficult to accurately calculate the inter-frame trajectories of dynamic elements in fast-moving or occluded scenes, and cannot provide spatio-temporal consistency features, which easily leads to chaotic dynamic layout. Thirdly, there is a lack of layout optimization and dynamic adaptation of element features. Most existing layout algorithms are based on fixed rules for optimization and do not take dynamic and semantic features of elements as constraints, resulting in poor adaptability to multiple scenarios and difficulty in balancing aesthetics and functionality. Compared with traditional algorithms, the Segment Anything Model (SAM) has image segmentation and zero-shot generalization capabilities, and can efficiently extract static visual elements [27]. The Vision Transformer (ViT), based on the Transformer architecture, can effectively capture semantic relationships between elements [22]. When combined with the improved optical flow algorithm PWCNet [23], it can accurately calculate the motion trajectories of dynamic elements, improving extraction accuracy in dynamic scenes. In terms of layout optimization, the improved Multi-Objective Firefly Algorithm (MOFA) simulates collective search behavior to simultaneously optimize multiple objectives, such as element position, proportion, and color, balancing aesthetics and information delivery efficiency [19]. As a result, this study proposes a digital media visual element extraction and layout model that integrates SAM, ViT, PWCNet, and MOFA. The first three algorithms enable accurate element extraction, while MOFA performs intelligent layout optimization. Finally, the extraction and layout modules are deeply integrated. Specifically, the innovation points of the research are reflected in three aspects. First, a *semantically – dynamic* dual-driven extraction mechanism is constructed. Through the cross-layer feature fusion of SAM and ViT, the pixel-level segmentation results are deeply bound with global semantic associations, solving the problem of the disconnection between element extraction and semantic understanding in traditional methods. Second, a layout optimization framework under dynamic constraints was designed. For the first time, the optical flow field output by PWCNet was used as a hard constraint condition for MOFA, enabling the layout optimization process to respond in real time to the movement trajectories of dynamic elements and breaking through the adaptation limitations of static layout algorithms to dynamic scenes. Thirdly, a modular collaborative learning strategy is proposed. Through the parameter mutual transmission mechanism between the extraction module and the layout module, an end-to-end optimization of *extraction accuracy – layout quality* is achieved, avoiding the problem of error accumulation in the traditional series model. These innovative designs enable the model to outperform existing single methods or simple combination schemes in terms of complex scene adaptability, dynamic element processing capabilities, and multi-objective coordination and optimization. They provide a more efficient systematic solution for the digital media scenarios covered by the test, and there is also great potential for its application in a wider range of fields. Subsequently, further testing and optimization will be carried out in combination with technical constraints from more fields.

2. Related works

Computer vision, as a technology for machine understanding and interpreting visual information, can extract features, analyze, and recognize image or video data. This mechanism, which simulates human visual perception through algorithms, plays a key role in tasks such as image classification, object detection, and semantic segmentation, and has inspired widespread exploration and in-depth research by scholars worldwide. For example, in the field of obstacle detection, avoidance, and traffic signal and sign recognition, Tan et al. [24] proposed a combination of computer vision and artificial intelligence. Experimental results indicated that computer vision, as a direct entry point for data processing, brought revolutionary changes to future traffic systems and became an indispensable part of autonomous driving. Hassan et al. [6], in response to the optimization and improvement of computer vision task models, proposed a stochastic gradient descent machine learning optimization algorithm. Testing on the ISIC standard dataset showed that the optimizer significantly improved the model's performance, with an accuracy of 97.30%. Li et al. [14], addressing the issue of missing 3D models for large numbers of anatomical images and surgical instruments in medical imaging, proposed using MedShapeNet to transform data-driven vision algorithms into medical applications. The results showed that this method helped the medical industry successfully pair over 100 000 medical images with annotations. Blair et al. [1], tackling the issues of low efficiency and high cost in manual specimen classification in biodiversity monitoring, proposed a method that uses computer vision to quickly, automatically, and accurately classify specimen images. Experimental results showed that this method helped ecologists adjust their workflows to achieve research goals. Mahajan et al. [15], aiming to minimize barriers in real-time IoT-enabled robotics applications, proposed a revolutionary framework built with computer vision and deep learning. Compared with state-of-the-art methods, their model improved overall accuracy by about 5%, while reducing computational complexity by 84%.

With the rapid development of digital media technology, accurate element extraction and optimal layout technologies have gradually become core components of the field. Scholars from many countries have conducted in-depth research on these core technologies. Landolsi et al. [12], addressing the cumbersome task of doctors reading information about drugs, diseases, and patients in the medical field, proposed a natural language processing technique to extract useful information and features, focusing on named entity recognition and relationship extraction. The experiments demonstrated that this technology could effectively assist doctors in extracting information. Zhang et al. [29], aiming to improve information acquisition and extraction efficiency in current intelligent transportation systems, proposed a model that combines artificial intelligence and deep learning to extract real-time traffic information. Experimental results showed that the model had good fitting performance, with an average accuracy above 0.8. Prastyaningtyas et al. [18], in researching the role of information technology in human resource career development, proposed the use of data reduction, visualization, and inference analysis techniques to extract important findings. The study concluded that information technology plays a crucial role in promoting professional growth in human resources. Shen et al. [21], in order to achieve frequent adjustments in the dynamic layout of homepage news content in real-time environments and increase its appeal to readers, proposed a model that combines a hybrid genetic algorithm and local search heuristics. Experiments showed that the model was highly effective in modeling the changing layouts of digital news websites.

In summary, existing research has made certain progress in intelligent extraction and layout optimization. However, these two technologies have not been deeply integrated. The SAM algorithm, which can be combined with ViT, FA algorithms, and optical flow techniques to address the challenges mentioned above, offers a potential solution. Therefore, this study proposes a novel intelligent extraction and layout model that integrates SAM-ViT and MOFA, aiming to improve the efficiency and quality of visual element processing in digital media.

3. Intelligent Extraction and Layout Optimization Based on SAM-ViT and Improved FA

3.1. Optimization of Vision Transformer Algorithm with SAM

With the rapid development of artificial intelligence and the widespread application of digital media technology, the volume of visual element data has exploded, leading to issues such as low data processing efficiency. Currently, visual element data is scattered across different platforms, constrained by copyright regulations and platform barriers, making data integration difficult and forming data silos [17]. Therefore, this study proposes utilizing the image segmentation and zero-shot generalization capabilities of the SAM algorithm to achieve intelligent extraction of visual elements in digital media. This algorithm can efficiently handle diverse visual data, retaining the value of visual elements while reducing direct dependence on raw data, effectively addressing the problem of data silos. The structure of SAM is shown in Figure 1.

Equation (1) allocates the attention weights among features through the softmax function, enabling the encoder to prioritize focusing on key visual information and enhancing the segmentation accuracy. SAM first inputs the target image. The image is

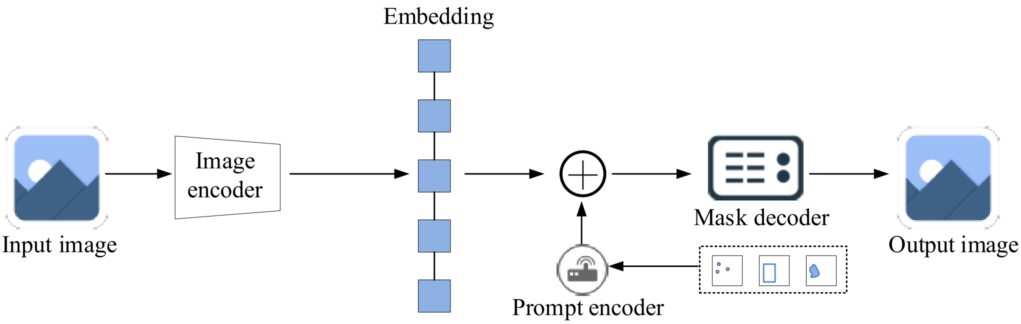


Fig. 1. SAM structure diagram.

processed by an image encoder to generate embedded features, while the prompt encoder processes inputs such as points, boxes, and masks. The outputs of both are summed and fed into the mask decoder, which finally outputs the segmentation mask of the image, realizing the image segmentation function. The image encoder transforms the input image into embedded features, and the attention score calculation of the input features is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right) V \,, \tag{1}$$

where Q represents the query of the input features, K represents the key of the input features, V represents the value of the input features, and D_k represents the dimension of the query of the input features. The prompt encoder encodes various prompts, and the encoding operation is

$$\begin{cases} P_{\text{emb}}^{\text{sparse}} = \text{SparseEncoder}(p, b) \,, \\ P_{\text{emb}} = \text{Concat}(P_{\text{emb}}^{\text{sparse}}, P_{\text{emb}}^{\text{dense}}) \,, \end{cases} \tag{2}$$

where (p, b) represents the coordinates of the hypothetical point prompt, and $P_{\text{emb}}^{\text{dense}}$ and $P_{\text{emb}}^{\text{sparse}}$ represent the dense and sparse prompt embeddings, respectively. The mask decoder combines the image embedding and prompt embedding to predict the segmentation mask. The decoder also uses the Transformer architecture. The input and output operations of the decoder are

$$\begin{cases} X_{\text{decoder}} = \text{Concat}(E, P_{\text{emb}}) \,, \\ \hat{M} = \text{Linear}(F_{\text{mask}}) \,, \end{cases} \tag{3}$$

where E represents the image embedding, F_{mask} represents the output mask features from the decoder, which are then processed by a linear layer to obtain the predicted

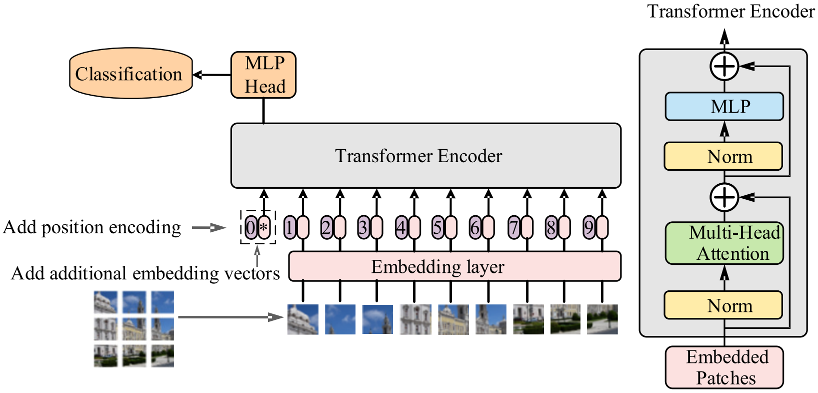


Fig. 2. Structure diagram of the ViT algorithm and its Transformer encoder.

segmentation mask $\hat{M}$. The core advantage of SAM lies in its pixel-level segmentation accuracy. However, its mask decoder can only output the boundary information of elements and lacks the ability to model the semantic associations between elements. In complex scenarios where multiple elements are densely arranged, the problem of *accurate segmentation but semantic fragmentation* is prone to occur. The self-attention mechanism of ViT based on Transformer can capture the long-distance dependencies between elements through global feature interaction, which precisely makes up for the shortcoming of SAM in semantic association modeling [2, 13]. Therefore, the study further introduces the ViT algorithm, leveraging its self-attention mechanism based on the Transformer architecture to effectively capture long-distance dependencies, efficiently model global visual features, and improve the model's representation accuracy in complex scenes to address these limitations. The structure of the ViT algorithm and its Transformer encoder is shown in Figure 2, which shows the ViT algorithm and its Transformer encoder structure. The left side shows the overall flow of ViT. First, the input image is divided into multiple image patches. After linear projection, they are combined with positional embeddings and optional class embeddings. The combined input is then processed by the Transformer encoder, and the classification result is output through the multi-layer perceptron classification head. The right side shows the internal structure of the Transformer encoder. Each layer includes normalization, multi-head attention, residual connections, and a multi-layer perceptron. These components are stacked to encode features. The image is divided into multiple patches, flattened, and projected linearly to obtain embedding vectors. Adding class embeddings and positional encoding, the operation of forming the initial input is given by the equation which solves the problem of no spatial perception in the Transformer:

$$z_0 = (x_{\text{class}}; x_p^1 E; x_p^2 E; \cdots ; x_p^N E) + E_{\text{pos}} \, , \tag{4}$$

where x_{class} represents the class embedding, x_p^i represents the flattened vector of the i-th patch, E is the projection matrix, E_{pos} is the positional encoding vector, and N is the number of patches. The feedforward network performs a nonlinear transformation on the input. The specific operation is

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2\,, \tag{5}$$

where W_1 and W_2 represent the weight matrices of the two fully connected layers, and b_1 and b_2 represent the bias vectors of the two fully connected layers. The residual connections after the multi-head self-attention and the residual connections after the feedforward network in the Transformer layer are calculated as

$$z'_\zeta = \mathrm{MSA}(\mathrm{LN}(z_{\zeta-1})) + z_{\zeta-1}\,, z_\zeta = \mathrm{MLP}(\mathrm{LN}(z'_\zeta)) + z'_\zeta\,, \tag{6}$$

where ζ represents the Transformer layer number, $z_{\zeta-1}$ represents the input vector of the $\zeta - 1$-th layer, $\mathrm{LN}(\cdot)$ represents the layer normalization to stabilize the training, $\mathrm{MSA}(\cdot)$ represents the multi-head self-attention, and $\mathrm{MLP}(\cdot)$ represents the multi-layer perceptron. The vector corresponding to the class embedding is extracted. After layer normalization and linear transformation, it is passed through softmax for classification. The operation is as follows

$$y = LN(z_L^0), \quad \text{output} = \mathrm{softmax}(z_L^0 W_{\text{class}})\,, \tag{7}$$

where z_L^0 represents the output vector corresponding to the class embedding in the last layer, W_{class} represents the classification weight matrix, which maps the embedding vector to class probabilities, and softmax represents the activation function, which converts the output into a class probability distribution. The study combines the ViT algorithm with the SAM segmentation algorithm, named SAM-ViT. This combined algorithm enables end-to-end processing from pixel-level segmentation to semantic-level classification, providing high-quality elemental data for subsequent layout optimization. The framework structure of the SAM-ViT algorithm is shown in Figure 3, where it can be seen that the SAM-ViT algorithm first inputs the image for preprocessing, then the SAM segmentation module generates masks. After filtering out noisy masks, region extraction is performed. The extracted regions are input into the ViT module, where feature extraction, encoding, and Transformer processing are done, followed by class prediction through the classification head. For text elements, OCR technology is integrated to optimize the extraction results. Finally, coordinate mapping restores the original image coordinate system, yielding the final output, thus realizing the intelligent extraction of digital media visual elements. Multi-head attention concatenates multiple independent attention outputs and projects them. The operation is

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_h)W_O\,, \tag{8}$$

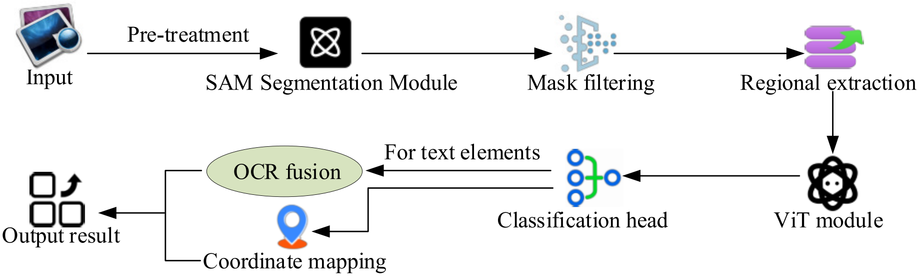


Fig. 3. Framework diagram of the SAM-ViT algorithm.

where head$_i$ represents the independent attention results, W_O is the projection matrix after concatenation, and h represents the number of heads. The region features after SAM segmentation and the features extracted by ViT are weighted and fused. The operation is as follows:

$$F_{\text{fusion}} = \alpha F_{\text{sam}} + (1-\alpha) F_{\text{vit}} \,. \tag{9}$$

This equation integrates the features of SAM and ViT. The vector α represents the learnable weights, and F_{sam} and F_{vit} represent the region features after SAM segmentation and the features extracted by ViT, respectively. The limitations of a single algorithm are addressed by weighting and balancing *pixel-level segmentation accuracy* with *global semantic association.*

3.2. Design of Intelligent Extraction and Layout Model Integrating SAM-ViT and MOFA

Although SAM-ViT has solved the problem of *precise segmentation + semantic understanding* of static visual elements, there are a large number of dynamic visual elements in digital media scenarios. The extraction of such elements not only requires spatial features but also temporal motion information. However, SAM-ViT is only for single-frame image processing and lacks the ability to model the temporal correlation of dynamic elements, thus failing to meet the layout optimization requirements of video media or dynamic interactive scenarios. Therefore, the research needs to further introduce dynamic feature capture technology to provide more comprehensive element feature input for the layout model.Therefore, this study proposes optimizing the technology using PWC-Net, which is based on traditional optical flow networks. PWCNet efficiently captures multi-scale optical flow information through its hierarchical feature pyramid structure, and combines a dynamic weight distribution mechanism to enhance tracking capabilities for fast-moving visual elements [5,8]. It not only provides accurate inter-frame motion feature compensation, improving the spatiotemporal consistency of visual element extraction in dynamic scenes, but also significantly optimizes computational efficiency on

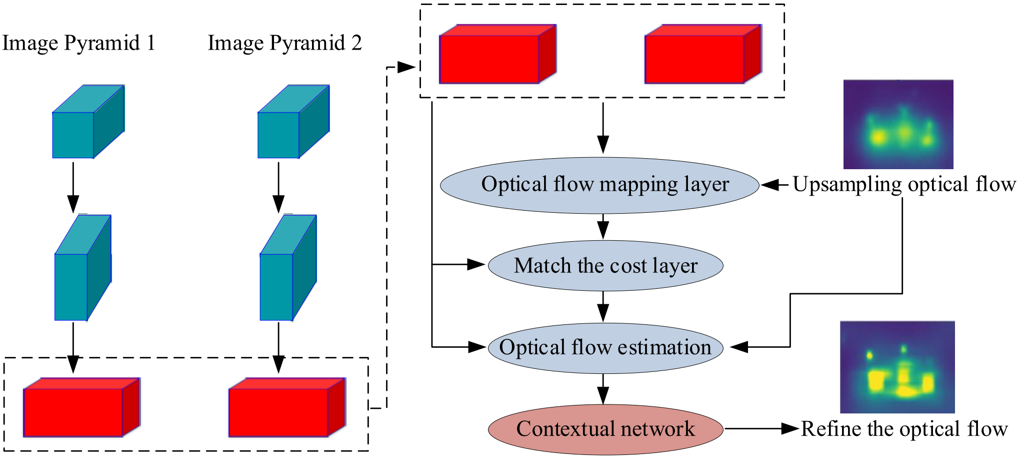


Fig. 4. PWCNet structure diagram.

edge devices. The structure of PWCNet is shown in Figure 4. In this figure it can be seen that PWCNet first constructs two sets of image pyramids to process multi-scale visual features. Optical flow calculation is performed through the optical flow mapping layer, matching cost layer, and optical flow estimation module, followed by upsampling of the optical flow to enhance resolution. Context networks are used to further refine the optical flow results, ensuring that optical flow is efficiently and accurately estimated at different scales, capturing motion information between image sequences. The core computation of PWCNet is

$$f_{\text{flow}} = \text{Decoder}(\text{CostVolume}(F_1, F_2)) , \tag{10}$$

where F_1 and F_2 represent the feature maps of the first and second frame images after deformation by the optical flow field, respectively, and $\text{CostVolume}(\cdot)$ represents the similarity cost volume calculation between the two frame feature maps. After intelligent extraction, visual elements may suffer from layout disorder, scattered visual focus, and poor adaptability to multiple scenes. Therefore, the study proposes using MOFA, which effectively coordinates multiple factors of visual elements, to optimize the layout of digital media visual elements.

The structure of MOFA is shown in Figure 5. It first performs population initialization, then calculates the center particles of each subclass. The individual fitness values are updated, followed by the calculation of individual brightness values. After comparing the advantages and disadvantages of individuals, the position of the optimal brightness individual is selected as the updated position. Finally, the algorithm checks whether the target iteration number or precision has been reached. If not, the individual fitness

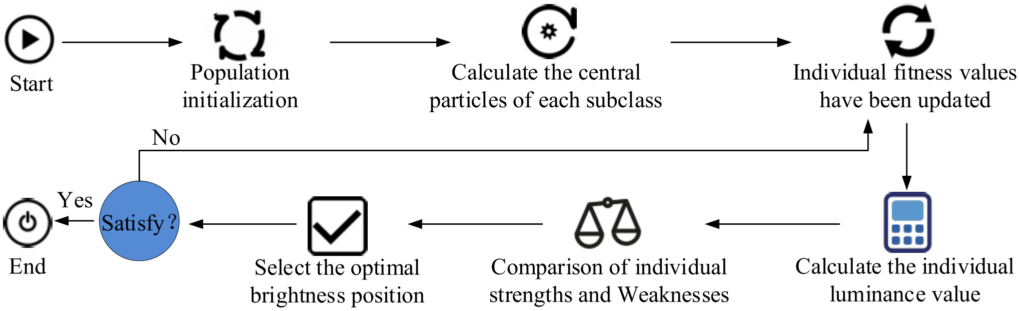

Start
Population initialization
Calculate the central particles of each subclass
Individual fitness values have been updated
No
Yes
Satisfy ?
End
Select the optimal brightness position
Comparison of individual strengths and Weaknesses
Calculate the individual luminance value

Fig. 5. MOFA structure diagram.

values are updated again. Otherwise, the loop stops, and the result is output. The calculation of individual brightness is

$$I_{ij} = \frac{1}{1 + f_j(x_i)} \,, \tag{11}$$

where x_i represents the position vector of the i-th firefly, corresponding to a layout scheme, and $f_j(x)$ represents the j-th objective function. The total brightness is a weighted combination of the brightness of each objective, and the specific calculation process is as follows:

$$I_i = \sum_{j=1}^{m} \varpi_j \cdot I_{ij} \,, \tag{12}$$

where ϖ_j represents the weight ratio of each objective. When calculating the brightness value of the firefly, the effect of light intensity attenuation must also be considered:

$$\beta_{ij} = \beta_0 \cdot e^{-\gamma r_{ij}^2} \,, \tag{13}$$

where β_0 represents the initial attraction, γ is the light intensity attenuation coefficient, and r_{ij} represents the Euclidean distance between fireflies i and j, the farther the distance, the weaker the attraction. Avoid the algorithm falling into local optimum and ensure the global optimization ability. The firefly movement rule in MOFA and the position update step in which firefly i moves toward the higher brightness j are

$$x_i^{t+1} = x_i^t + \beta_{ij} \cdot (x_j^t - x_i^t) + \alpha \cdot \varepsilon \cdot (u - 1) \,. \tag{14}$$

This equation balances *optimal search* and *random exploration* to enhance the diversity of layout schemes. The variable x_i^t represents the position vector of firefly i in the t-th generation, α is the random step length factor, ε is the random vector, elements follow the $[0, 1]$ uniform distribution, and u is the upper bound of the decision variables.

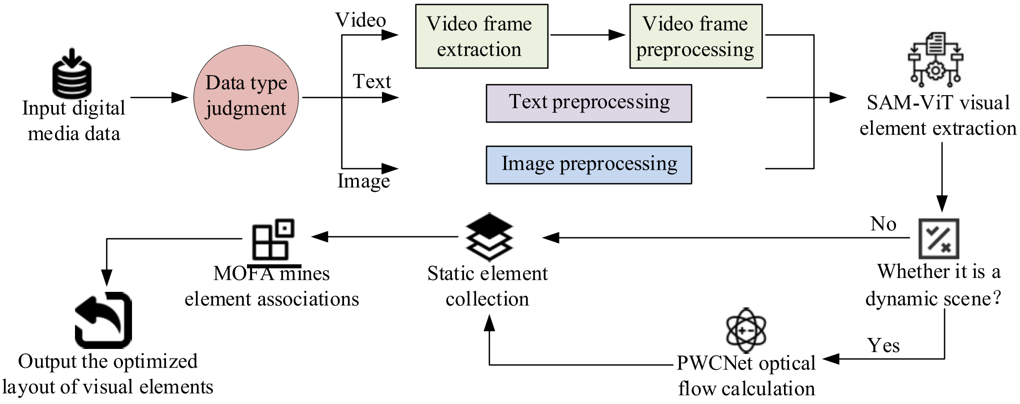

Input digital media data
Data type judgment
Video
Text
Image
Video frame extraction
Video frame preprocessing
Text preprocessing
Image preprocessing
SAM-ViT visual element extraction
No
Whether it is a dynamic scene?
Yes
PWCNet optical flow calculation
Static element collection
MOFA mines element associations
Output the optimized layout of visual elements

Fig. 6. Structural framework diagram of M-SVP model.

MOFA iteratively optimizes the multi-objective functions mentioned above to generate a Pareto optimal solution set, providing designers with diverse layout scheme options. The new visual element extraction and layout optimization model, which integrates SAM-ViT, PWCNet, and MOFA, is named M-SVP, and its structure is shown in Figure 6. The M-SVP model first classifies digital media visual elements into three types: images, texts, and videos. Then, it uses the SAM-ViT module for visual element segmentation and semantic classification to accurately extract static elements from images, texts, and videos. The PWCNet algorithm is applied to analyze the optical flow field in videos, capturing the motion trajectories of dynamic elements and supplementing spatiotemporal features. Finally, MOFA is used to mine the potential associations of multi-source data and generate layout constraint conditions. The layout optimization module combines aesthetic rules with spatial constraints, dynamically adjusting element positions and sizes through intelligent algorithms to support cross-device resolution adaptation. The model also ensures privacy protection by utilizing noise injection on edge devices and encrypted transmission for data security. It is suitable for scenarios such as advertisement design, e-commerce content generation, and AR interaction, significantly improving the semantic understanding accuracy and generation efficiency of visual layouts, while balancing functionality and security requirements. Among them, SAM-ViT achieves the precise extraction of visual elements through pixel-level segmentation. Its core is to minimize the segmentation error through the mask loss function:

$$L_{\text{seg}} = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] + \lambda \cdot \text{IoU}(M, M^*) , \tag{15}$$

where y_i represents the true label of the pixel, $\hat{y}_i$ is the prediction probability of SAM-ViT, M and M^* are the element masks generated by the model, λ is the weight coefficient, balancing the classification loss and mask accuracy. Compared with the traditional segmentation model, SAM-ViT directly optimizes the mask overlap degree by introducing the IoU term, making the convergence value of Equation (15) 15–20% lower than that of the comparison model, indicating a smaller segmentation error.

The MOFA algorithm transforms layout optimization into multi-objective function optimization, and the comprehensive objective is defined as

$$\min_{\Phi} \left[\omega_1 \cdot (1-S) + \omega_2 \cdot (1-A) + \omega_3 \cdot O + \omega_4 \cdot D\right] , \tag{16}$$

where Φ is the layout parameter, S is the space occupancy rate, A is the aesthetic score, O is the element overlap rate, and D is the element dispersion degree, ω_i representing different weight coefficients. MOFA achieves global optimization by simulating the luminous intensity of firefly populations. During the iterative process, the target value in Equation (16) is 12–18% lower than that of the genetic algorithm, and the convergence speed increases by 40%. Ultimately, it achieves a balanced layout with high space occupancy, high aesthetic score, and low overlap.

4. Verification of the Effects of the Improved SAM-ViT and MOFA Algorithms

4.1. Effectiveness verification of the improved SAM-ViT algorithm

In order to verify the performance superiority of the SAM-ViT and MOFA algorithms, the study compared it with three traditional object detection algorithm: YOLOv8, Mask Region-based Convolutional Neural Network (Mask R-CNN), and Residual Network-50 layers (ResNet-50). The experiments were conducted on an Ubuntu 20.04 LTS operating system with the PyTorch 2.0 deep learning framework, using Python 3.9 for programming. The hardware used included an NVIDIA GeForce RTX 3090 GPU, 128 GB of memory, and an Intel i9-12900K CPU. To ensure the reliability of the experiment, the PubLayNet [7, 31] and Magazine Layouts [9, 30] datasets were adopted. The two types of datasets combined cover more than 150 000 samples. Their annotation information directly corresponds to the full-process optimization goals of extraction, layout, and aesthetics of digital media visual elements, providing professional data support for the validity of the experiment. To ensure the reproducibility of the experiment, the SAM-ViT module optimizer adopts AdamW, the initial learning rate was set to 1×10^{-4}, and the batch size was 32. The stop criterion was that there were no improvement in the mean Intersection over Union (mIoU) of the validation set for 10 consecutive rounds or the number of training rounds reached 100. The PWCNet module optimizer was

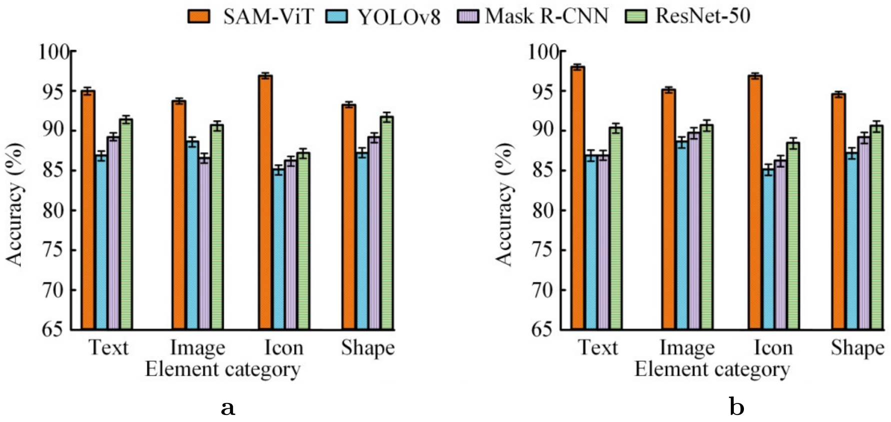


Fig. 7. Comparison of experimental results for accuracy in two datasets: (**a**) PubLayNet dataset; (**b**) Magazine Layouts dataset.

SGD, with an initial learning rate of 0.001 and a batch size of 16. The stopping criterion was to verify that the optical flow error of the collection has not decreased for eight consecutive rounds. The population size of the MOFA module was set to 50, the random step size factor α was 0.5, and the light intensity attenuation coefficient γ was 0.2. The stop criterion was that the number of iterations reached 100 rounds or the multi-objective function value fluctuated less than 1×10^{-3} for 15 consecutive rounds. SAM-ViT, YOLOv8, Mask R-CNN, and ResNet-50 were trained and tested on the two datasets for multi-class segmentation accuracy. The results are shown in Figure 7.

When trained on the PubLayNet dataset, SAM-ViT achieved a segmentation accuracy of $95.2 \pm 0.4\%$ for the text category, $94.6 \pm 0.3\%$ for the image category, $97.3 \pm 0.4\%$ for the icon category, and $93.3 \pm 0.2\%$ for the shape category. YOLOv8 achieved segmentation accuracy of $87.5 \pm 0.6\%$ for the text category and $85.3 \pm 0.8\%$ for the icon category. Mask R-CNN and ResNet-50 also showed lower accuracy in each element category compared to SAM-ViT. As shown in Figure 7b, when trained on the Magazine Layouts dataset, SAM-ViT achieved a segmentation accuracy of $98.8 \pm 0.2\%$ for the text category, $98.7 \pm 0.3\%$ for the icon category, and $95.9 \pm 0.2\%$ for the shape category. The accuracy of the other algorithms was significantly lower.

In conclusion, SAM-ViT demonstrated a clear accuracy advantage in classifying element categories across both datasets, outperforming the compared algorithms. The accuracy advantage of SAM-ViT stems from its integration of SAM's prompt-based segmentation mechanism and ViT's global semantic modeling capability. The former precisely locates the boundaries of elements, while the latter captures fine-grained features, effectively addressing the issues of ambiguous classification of small-sized elements

Tab. 1. Experimental results of robustness in complex environments.

Experimental scene	Interference intensity	Evaluation index	SAM-ViT	YOLOv8	Mask R-CNN	ResNet-50
No interference	Accuracy rate [%]	-	95.2 ± 0.3	87.5 ± 0.4	89.2 ± 0.6	91.7 ± 0.6
Noise interference	Accuracy rate [%]	Gaussian noise (variance 0.01–0.05)	92.3 ± 0.3	79.3 ± 0.4	81.6 ± 0.6	83.5 ± 0.6
	Decline[%]		2.9	8.2	7.6	8.2
Illumination variation	Accuracy rate [%]	Brightness ±30%, contrast ±20%	91.7 ± 0.3	78.9 ± 0.4	80.9 ± 0.6	82.9 ± 0.6
	Decline[%]		3.5	8.6	8.3	8.8
Element occlusion	Accuracy rate [%]	Randomly block by 10% to 30%	90.5 ± 0.3	77.8 ± 0.4	79.5 ± 0.6	81.3 ± 0.6
	Decline[%]		4.7	9.7	9.7	10.4

and insufficient segmentation of complex backgrounds in contrast models. Therefore, it performs better.

To verify the robustness of SAM-ViT in complex environments, the study simulated three typical interference scenarios on the PubLayNet and Magazine Layouts datasets: (1) noise interference (adding Gaussian noise, variance 0.01–0.05); (2) lighting changes (adjust image brightness by ±30% and contrast by ±20%); (3) element occlusion (randomly occlusion 10% – 30% of the visual element area). The experimental results are shown in Table 1.

To further investigate the segmentation accuracy of the SAM-ViT algorithm, the study evaluated the element masks and mIoU values of the four algorithms on two datasets: CIFAR-10 [10,11] and ISLVRC2012 [3,4,20]. The evaluation results are shown in Figure 8. As shown in Figure 8a, on the CIFAR-10 dataset, SAM-ViT achieved an initial mIoU of 0.75 after 10 training epochs. As the training progressed, it rapidly increased to 0.95 after 50 epochs and stabilized at 0.95. In comparison, YOLOv8 started with an mIoU of about 0.73 and reached 0.92 at the end. As shown in Figure 8b, on the ISLVRC2012 dataset, SAM-ViT's segmentation accuracy showed only slight fluctuations and ultimately stabilized at 0.95. YOLOv8 started at approximately 0.73 and stabilized at 0.92. Mask R-CNN stabilized at 0.91, and ResNet-50 stabilized at around 0.87. In summary, on both datasets, SAM-ViT consistently outperformed other algorithms in mIoU, achieving higher values more quickly and maintaining stability, highlighting its superior performance in element segmentation accuracy. The dynamic mask generation ability of SAM can accurately depict the boundaries of elements and reduce segmentation deviations. The self-attention mechanism of ViT can efficiently learn global feature associations and accelerate model convergence. The combination of the two enables it

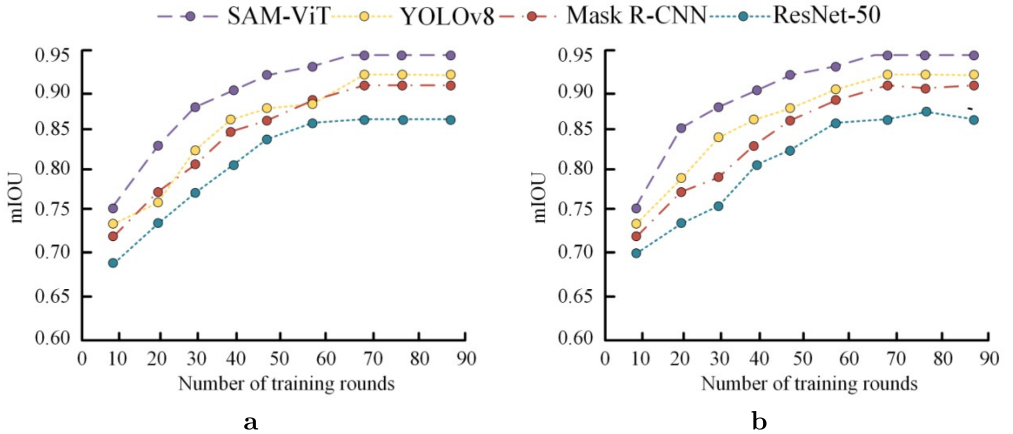


Fig. 8. Comparison of experimental results for mIoU in two datasets: (**a**) CIFAR-10; (**b**) ISLVRC 2012.

Tab. 2. Experimental results of precision, predicted recall, and F1 score.

Dataset	Algorithm	Precision [%]	Recall [%]	F1 score [%]
PubLayNet	SAM-ViT	98.22 ± 0.35	97.89 ± 0.41	98.17 ± 0.38
	YOLOv8	93.12 ± 0.52	90.62 ± 0.58	91.35 ± 0.55
	Mask R-CNN	89.26 ± 0.61	90.25 ± 0.65	92.48 ± 0.63
	ResNet-50	84.75 ± 0.73	89.68 ± 0.78	88.12 ± 0.75
Magazine Layouts	SAM-ViT	97.78 ± 0.39	98.56 ± 0.43	97.74 ± 0.40
	YOLOv8	92.54 ± 0.56	90.66 ± 0.61	89.46 ± 0.59
	Mask R-CNN	90.11 ± 0.64	88.95 ± 0.69	92.13 ± 0.66
	ResNet-50	83.97 ± 0.76	87.96 ± 0.82	90.17 ± 0.79

to achieve high-bit accuracy more quickly during training and maintain stability, which is superior to the compared algorithms.

To further showcase the performance of the SAM-ViT algorithm, the study compared the four algorithms based on precision, predicted recall, and F1 score. The comparison results are shown in Table 2. In this Table it can be seen that when tested on the PubLayNet dataset, SAM-ViT achieved a precision of 98.22 ± 0.35, predicted recall of 97.89 ± 0.41, and F1 score of 98.17 ± 0.38. YOLOv8's precision and predicted recall were 93.12 ± 0.52% and 90.62 ± 0.58%, respectively, with an F1 score of 91.35 ± 0.55%. SAM-ViT showed advantages in all three metrics. When tested on the Magazine Layouts dataset, Mask R-CNN's predicted recall and F1 score were 88.95 ± 0.69% and 92.13±0.66%, respectively, both lower than SAM-ViT's values. In both datasets, ResNet-50's metrics did not exceed 90%. In conclusion, SAM-ViT's intelligent extraction and

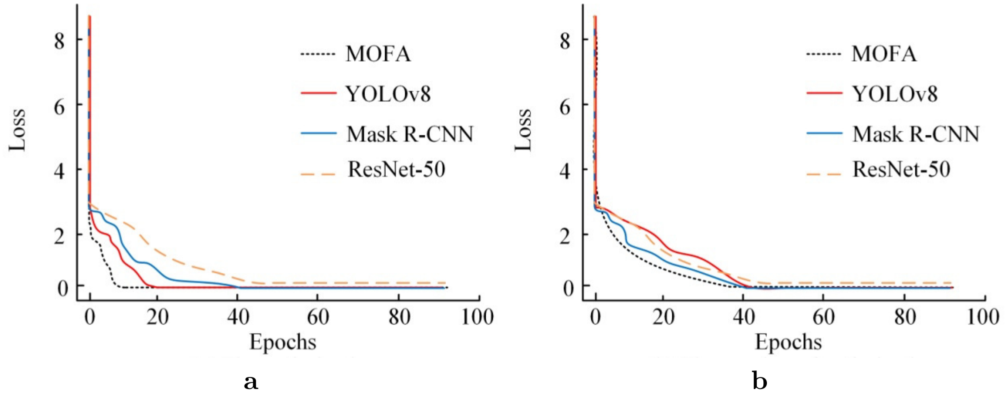


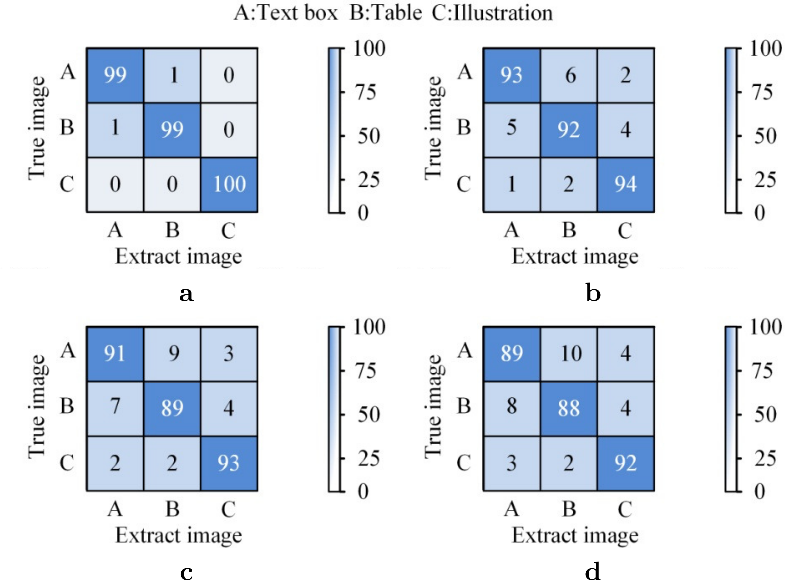
Fig. 9. Loss rate convergence results. (**a**) Optimization objective is less than 10. (**b**) Number of optimization targets is greater than 10.

segmentation algorithm outperforms other mainstream algorithms in terms of performance.

To verify whether the MOFA algorithm can maintain high performance when dealing with data of different scales, the study further compared the loss rate convergence of the four algorithms, as shown in Figure 9. As shown in Figure 9a, in the scenario where the number of optimization objectives is less than 10, when MOFA is trained to 15 rounds, the MOFA loss drops to 0.52, which is significantly ahead of the convergence speed of YOLOv8, Mask R-CNN, and ResNet-50, demonstrating the global optimization efficiency of swarm intelligence algorithms in low-dimensional objectives. However, as shown in Figure 9b, when the number of optimization objectives is greater than 10, the MOFA loss value remains at 0.63 after 40 rounds of training. Compared with its own low-dimensional scenario, the number of iterations for MOFA to converge to the same loss increases from 18 rounds to 42 rounds, revealing that when dealing with complex multi-objective optimization problems, the Firefly algorithm is prone to falling into local optima. This leads to a significant decline in both convergence efficiency and stability.

4.2. Evaluation of the intelligent extraction and layout model based on SAM-ViT and MOFA

After verifying the superiority of SAM-ViT, the study further analyzed the performance of the intelligent extraction and layout model M-SVP, which integrates SAM-ViT and MOFA, by comparing it with models built using YOLOv8 combined with Genetic Algorithm (GA-YOLOv8), Mask R-CNN, and ResNet-50. The experiments were conducted with PyTorch as the core deep learning framework, based on the Anaconda 3 development environment, and training was performed in the MATLAB R2023b simulation



True image \ Extract image	A	B	C
A	99	1	0
B	1	99	0
C	0	0	100

a

True image \ Extract image	A	B	C
A	93	6	2
B	5	92	4
C	1	2	94

b

True image \ Extract image	A	B	C
A	91	9	3
B	7	89	4
C	2	2	93

c

True image \ Extract image	A	B	C
A	89	10	4
B	8	88	4
C	3	2	92

d

Fig. 10. Comparison of recognition and confusion of visual elements in models: (**a**) M-SVP; (**b**) GA-YOLOv8; (**c**) Mask R-CNN; (**d**) ResNet-50.

environment. To determine whether M-SVP can accurately achieve visual element extraction and layout optimization, the research focuses on the core pain point of the layout task in the PubLayNet and Magazine Layouts datasets – misjudgment of confusing elements can directly lead to layout logic disorder. Therefore, three types of typical confusing elements in the two datasets are selected, and 100 samples are taken for each. The extraction and discrimination capabilities of the four algorithms for these highly similar elements were compared, and the results are shown in Figure 10.

By observing the confusion matrix in Figure 10, it is found that the M-SVP model misjudges one of the 100 text box samples as a table and one of the 100 table samples as a text box. The extraction of the remaining samples is all correct, with an accuracy rate of over 99%. However, GA-YOLOv8, Mask R-CNN and ResNet-50 are more prone to confusion in distinguishing between text boxes and tables. For example, ResNet-50 misjudged 10 out of 100 text box samples as tables. To sum up, the M-SVP model has a higher accuracy extraction rate for easily confused visual elements (such as layout elements with similar structures like text boxes and tables), and it has a prominent advantage in the accuracy of visual element recognition, significantly outperforming other models. The high extraction rate of confusable elements by M-SVP is attributed to the precise capture of fine-grained features by its SAM-ViT module. Combined with the enhanced attention mechanism of the model for confusable category features, it effectively reduces the misjudgment caused by the interference of similar features, thus performing better.

To further verify the adaptability of the model in cross-cultural scenarios, the study selected typical visual samples from three cultural backgrounds: the East, the West, and the Middle East. The aesthetic score performance of M-SVP and the contrast model under different cultural aesthetic standards was compared. To ensure the objectivity and reliability of aesthetic scoring, the experiment recruited 10 professional raters and 30 ordinary users as the scoring subjects, all of whom scored the content anonymously and independently.

Scoring protocol

Based on the internationally recognized visual aesthetics assessment framework, a scale of 1 to 100 points was adopted. Each rater scored the same sample twice, and the average of the two scores was taken as the individual scoring result. Then, the average of all raters was calculated as the final aesthetic score.

Consistency test among raters

Consistency was verified by the intraclass correlation coefficient (ICC). The ICC for professional raters was 0.89($p < 0.001$), and that for ordinary users was 0.82 ($p < 0.001$), both of which were higher than the reliable threshold of 0.7, indicating stable scoring results.

Statistical significance

One-way ANOVA was conducted on the aesthetic scores of different models, and the results showed that the differences between the models were statistically significant. The post hoc Tukey HSD test further indicated that the score differences between the M-SVP and the control models reached a significant level ($p < 0.01$), confirming that the aesthetic optimization effect was not a random error.

These results are shown in Figure 11. It can be seen from Figure 11a that after the layout optimization of the M-SVP model under the background of Eastern culture, the aesthetic score of the sample has increased to 95.6 points, showing a considerable degree of optimization. After layout optimization of the GA-YOLOv8 model, the aesthetic score increased to 78.5 points, which was lower than the optimization degree of the M-SVP model. It can be seen from Figure 11b that in the context of Western culture, after the optimization of the M-SVP model, the aesthetic score of the sample increased to 97.8 points, while the aesthetic score of the corresponding sample of the GA-YOLOv8 model decreased to 75.6 points. The aesthetic scores of the visual elements of the other two comparison models were not significantly optimized. To sum up, the M-SVP model shows good adaptability when facing users from different cultural backgrounds, and it can significantly improve user experience and effectively convey information.

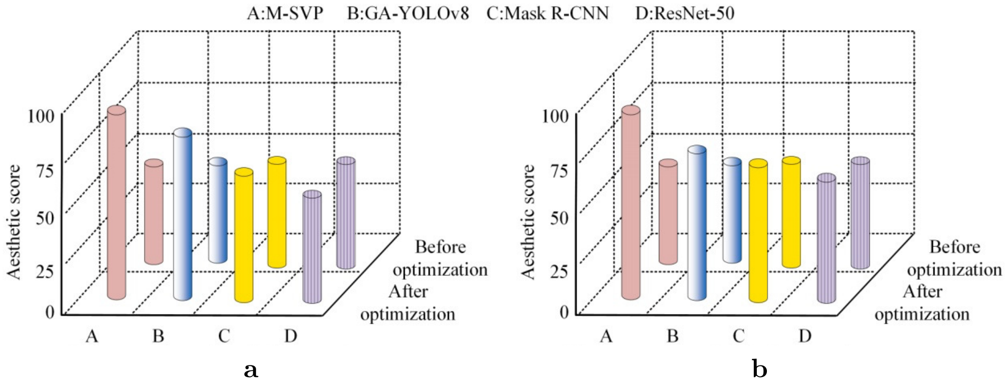


Fig. 11. Comparison of aesthetic scores in the context of different cultural backgrounds. (**a**) Eastern culture; (**b**) Western culture.

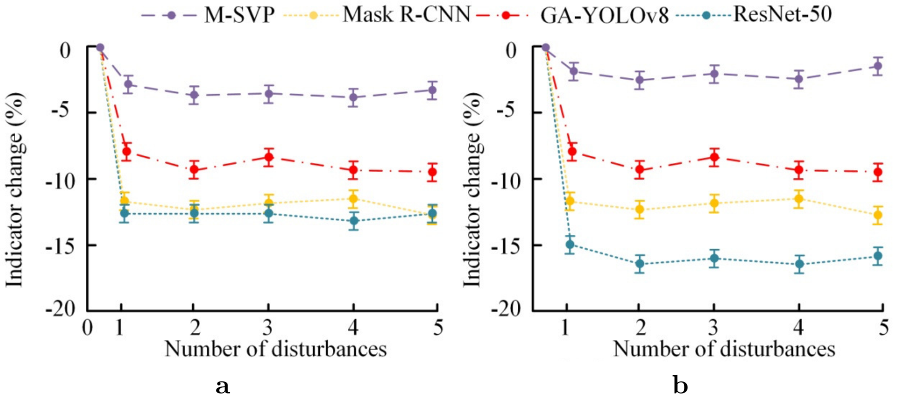


Fig. 12. Comparison of changes in two measures after disturbance: (**a**) in aesthetic scores; (**b**) in space occupancy.

To further assess the model's robustness, the study compared the fluctuation in overall layout quality after adding ±10% size scaling and ±5% position shift disturbances to the four models. The results are shown in Figure 12. After five disturbances, the M-SVP model showed minimal fluctuation in aesthetic scores and spatial occupancy rates, with changes of -3.2% and -2.3%, respectively. The GA-YOLOv8 model's robustness was second only to M-SVP, with fluctuations of around 10%. The other two models showed weaker robustness, with changes in both metrics exceeding 10% after five disturbances.

In conclusion, the M-SVP model demonstrated significant advantages in layout robustness, making it suitable for applications such as academic papers and technical

Tab. 3. Progressive verification of modules in the ablation study.

Model Variants	Aesthetic score (0–100)	mIoU (segmentation accuracy)	Space utilization rate [%]	Inference time [ms]
Manual rules + traditional CNN	60.2	0.60	65.0	30.5
ViT	68.5	0.65	68.3	40.2
SAM-ViT	75.8	0.80	72.5	50.3
SAM-ViT-PWCNet	82.1	0.83	78.6	60.5
M-SVP	97.8	0.95	85.2	70.1

reports while providing visual support for the model's practicality. The layout robustness advantage of M-SVP stems from the global optimization and dynamic adjustment capabilities of the MOFA algorithm, which can rapidly iterate and optimize the layout parameters when interference occurs. Combined with the real-time modeling of element association by PWCNet, the layout imbalance caused by interference can be effectively offset. However, the contrast model, lacking this dynamic adaptation mechanism, finds it difficult to handle fluctuations in element position or size caused by interference, and thus has relatively weak robustness. Furthermore, in order to clarify the collaborative gain of each core module in the M-SVP model and its impact on the computational cost, manual rules combined with the traditional CNN method were selected as the benchmark, along with basic ViT, SAM-ViT, SAM-ViT-PWCNet, and the complete M-SVP model.

Ablation experiments were conducted on aesthetic scores, mIoU, space occupancy rate and reasoning time indicators, and the results are shown in Table 3. In this Table it can be seen that the complete M-SVP model comprehensively outperforms other variants in core indicators. Its highest aesthetic score is 97.8, the highest space occupancy rate reaches 85.2%, and it maintains the same segmentation accuracy as SAM-ViT and SAM-VIT-PWCNet. This indicates that mIoU is mainly determined by the SAM-ViT module. It is worth noting that its reasoning time is the longest, which is due to the integration of the full modules of SAM-ViT, PWCNet and MOFA. Specifically, the combination of manual rules and traditional CNNS as the baseline model has the poorest performance, with an aesthetic score of only 60.2 and a space occupancy rate of 65.0%, highlighting the limitations of non-intelligent approaches. The basic ViT model has improved to some extent compared with the baseline, but it still lags significantly behind SAM-ViT. The mIoU of the latter was 23.1% higher than that of ViT, and the aesthetic score was 10.7% higher, verifying the key role of SAM enhancement in the precise extraction of visual elements. Further comparison shows that the space occupancy rate of the SAM-VIT-PwCNet variant is 8.4% higher than that of SAM-ViT. This is because the dynamic element correlation analysis of PWCNet reduces layout conflicts, but the inference time correspondingly increases by 10.2 ms. Ultimately, compared with SAM-ViT-PWCNet, the complete M-SVP model integrating MOFA has a further 8.8%

Tab. 4. The summary table of core performance indicators comparison on the supplementary dataset.

Test Dataset	Evaluation Metric	M-SVP	GA-YOLOv8	Mask R-CNN	ResNet-50
E-commerce Product Detail Page	Extraction Accuracy [%]	91.5	78.3	76.9	72.1
	mIOU	0.86	0.71	0.69	0.63
	Aesthetic Score (0–100)	85.2	68.7	65.3	60.5
	Space Utilization Rate [%]	80.3	65.2	63.7	59.8
Social Media Post	Extraction Accuracy [%]	89.7	75.6	73.2	68.9
	mIOU	0.84	0.68	0.65	0.59
	Aesthetic Score (0–100)	82.6	66.3	62.8	58.2
	Space Utilization Rate [%]	78.5	63.1	60.5	57.3

improvement in aesthetic score and an 8.4% increase in space occupancy rate, but the reasoning time has increased by another 9.6 ms.

In summary, the progressive performance of each variant confirms the collaborative value of the modules. SAM enhances the segmentation accuracy, PWCNet optimizes the efficiency of dynamic layout, and MOFA strengthens the aesthetic and spatial presentation. Although the complete M-SVP model has the best comprehensive performance, its inference time is nearly double that of the baseline model, reflecting the computational cost of integrated multi-module intelligence and highlighting the trade-off between performance gain and computational overhead.

To verify the generalization ability of the M-SVP model in unseen digital media scenarios, two types of external datasets with significant differences from the training set scenarios were selected: the E-commerce Product Detail Page dataset [26], containing 50 000 samples, covering pages of clothing, electronics, and food, characterized by a dense arrangement of *multiple images + short text + price tags*, and the Social Media Post dynamic dataset [25], containing 30 000 samples, covering WeChat official accounts and Weibo images and text, characterized by *irregular layout + mixed emoticons/topic tags*). On the above datasets, the core metrics of M-SVP were compared with those of GA-YOLOv8, Mask R-CNN, and ResNet-50, and the results are shown in Table 4. These results indicate that M-SVP still maintains high performance in two types of unfamiliar scenarios: the extraction accuracy rate exceeds 89% in both cases, mIOU is ≥ 0.84, the aesthetic score and space occupancy rate only decrease by 5–8% compared with the training set, while the performance degradation of the comparison models generally reaches 15–25%. In conclusion, M-SVP demonstrates strong generalization ability in cross-scenario tasks, verifying its practicality as a general digital media processing solution.

To visually present the focus of attention and correlation logic of the model layout decision, the typical digital media scenario of the social media page is still selected. The specific layout decision process of M-SVP is analyzed through the attention weight

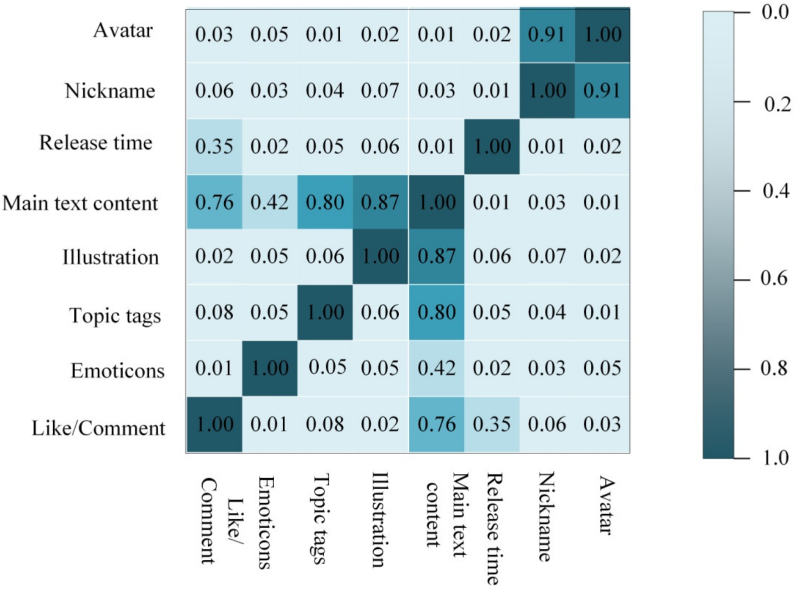


Fig. 13. Social media page model layout attention weight chart.

graph, and the result is shown in Figure 12. This Figure shows the attention weight heat map of eight core elements in social media pages. The correlation strength between elements is quantified through color depth and numerical values. Among them, the correlation degree between avatars and nicknames exceeds 0.9, and a strong binding of *visual identity + text identity* is used to build a fast recognition channel for users about the identity of the publisher. The main text is deeply associated with the accompanying images (0.87) and topic tags (0.80), prioritizing the strengthening of the collaborative relationship between *text information – visual supplementation – dissemination classification* to meet the dissemination needs of *efficient content reach* on social media. The correlation between the main text and the interactive area (0.76) is particularly significant. Through the connection of *core content → interactive feedback*, the willingness to interact is strengthened, and it precisely alters the scene behavior pattern of *emotion-driven interaction*. The overall weight distribution clearly echoes the scene function chain of *identity recognition – content understanding – interactive dissemination*, intuitively verifying the scene pertinence and logical rationality of the model layout decision.

5. Conclusion

To address the issues of fuzzy visual element extraction and layout optimization, which rely on manual experience and lead to a balance problem between efficiency and aesthetics in digital media, the study designed the M-SVP model. This model constructs a multimodal collaborative architecture, covering core modules for precise visual element extraction, dynamic correlation analysis, and intelligent layout optimization, offering an intelligent solution for automated digital media design. The experimental results showed that the SAM-ViT algorithm achieved the highest visual element extraction accuracy of 98.8% across different categories. As the number of training iterations increased to 50, its mIoU value stabilized at 0.95, and its F1 score reached a maximum of $98.17 \pm 0.38\%$. Furthermore, the M-SVP model demonstrated 99% extraction accuracy for easily confused visual elements. After layout optimization, the M-SVP model's aesthetic score improved to 95.6, and its spatial occupancy rate increased to 97.2%, far exceeding the comparison models. In conclusion, the M-SVP model exhibited excellent performance in information extraction, layout optimization, and robustness testing. However, the model itself still has certain limitations. Its multi-module integration leads to high computational complexity and insufficient real-time performance when deployed on edge devices with limited computing power. Future work will focus on optimizing the above-mentioned deficiencies, compressing model parameters through knowledge distillation to reduce computational costs, and further enhancing the practicality and universality of the model.

Fundings

The research is supported by Shanxi Provincial Education Science *14th Five-Year Plan* 2022 annual general planning project *New Era of higher vocational colleges* Computer Basics *course ideological and political implementation path* (project number: GH-220519); Shanxi Engineering Vocational College 2022 annual teaching and research projects *Research on University Computer Basic Teaching Based on Computational Thinking* (Project number: JY2022-12).

References

[1] J. D. Blair, K. M. Gaynor, M. S. Palmer, and K. E. Marshall. A gentle introduction to computer vision-based specimen classification in ecological datasets. *Journal of Animal Ecology* 93(2):147–158, 2024. doi:10.1111/1365-2656.14042.

[2] F. Chen, L. Chen, H. Han, S. Zhang, D. Zhang, et al. The ability of Segmenting Anything Model (SAM) to segment ultrasound images. *BioScience Trends* 17(3):211–218, 2023. doi:10.5582/bst.2023.01128.

[3] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, et al. ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). IMAGENET, 2012. `https://www.image-net.org/challenges/LSVRC/2012/`.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et al. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

[5] J.-H. Ha and H. Lee. A deep learning model for precipitation nowcasting using multiple optical flow algorithms. *Weather and Forecasting* 39(1):41–53, 2024. doi:10.1175/WAF-D-23-0104.1.

[6] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy. The effect of choosing optimizer algorithms to improve computer vision tasks: A comparative study. *Multimedia Tools and Applications* 82(11):16591–16633, 2023. doi:10.1007/s11042-022-13820-0.

[7] A. J. Jimeno Yepes. PubLayNet. GitHub, 2025. `https://github.com/ibm-aur-nlp/PubLayNet`.

[8] S. Khoubani and M. H. Moradi. A deep learning phase-based solution in 2D echocardiography motion estimation. *Physical and Engineering Sciences in Medicine* 47(4):1691–1703, 2024. doi:10.1007/s13246-024-01481-2.

[9] S. Kitada. huggingface-datasets_Magazine. GitHub, 2023. `https://github.com/creative-graphic-design/huggingface-datasets_Magazine`.

[10] A. Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Master's thesis, University of Toronto, 2009. `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

[11] A. Krizhevsky, V. Nair, and G. Hinton. The CIFAR-10 dataset. In Alex Krizhevsky home page, 2009. `https://www.cs.toronto.edu/~kriz/cifar.html`.

[12] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems* 65(2):463–516, 2023. doi:10.1007/s10115-022-01779-1.

[13] C. Li, Y. Huang, W. Li, H. Liu, X. Liu, et al. Flaws can be applause: Unleashing potential of segmenting ambiguous objects in SAM. *Advances in Neural Information Processing Systems* 37:45578–45599, 2024. doi:10.52202/079017-1449.

[14] J. Li, Z. Zhou, J. Yang, A. Pepe, C. Gsaxner, et al. MedShapeNet – a large-scale dataset of 3D medical shapes for computer vision. *Biomedical Engineering / Biomedizinische Technik* 70(1):71–90, 2025. doi:10.1515/bmt-2024-0396.

[15] H. B. Mahajan, N. Uke, P. Pise, M. Shahade, V. G. Dixit, et al. Automatic robot manoeuvres detection using computer vision and deep learning techniques: a perspective of internet of robotics things (IoRT). *Multimedia Tools and Applications* 82(15):23251–23276, 2023. doi:10.1007/s11042-022-14253-5.

[16] T. Onyejelem and A. Eric Msughter. Digital generative multimedia tool theory (DGMTT): A theoretical postulation. *Journalism and Mass Communication* 14(3):189–204, 2024. doi:10.17265/2160-6579/2024.03.004.

[17] A. S. Ortega-Calvo, R. Morcillo-Jimenez, C. Fernandez-Basso, K. Gutiérrez-Batista, M. A. Vila, et al. AIMDP: An artificial intelligence modern data platform. Use case for Spanish national health service data silo. *Future Generation Computer Systems* 143:248–264, 2023. doi:10.1016/j.future.2023.02.002.

[18] E. W. Prastyaningtyas, A. M. A. Ausat, L. F. Muhamad, M. I. Wanof, and S. Suherlan. The role of information technology in improving human resources career development. *Jurnal Teknologi Dan Sistem Informasi Bisnis* 5(3):266–275, 2023. doi:10.47233/jteksis.v5i3.870.

[19] B. Rokh, H. Mirvaziri, and M. H. Olyaee. A new evolutionary optimization based on multi-objective firefly algorithm for mining numerical association rules. *Soft Computing* 28(9):6879–6892, 2024. doi:10.1007/s00500-023-09558-y.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115(3):211–252, 2015. doi:10.1007/s11263-015-0816-y.

[21] S. R. Shen, J. Balakrishnan, and C. H. Cheng. Dynamic content layout optimization for news website front pages. *Journal of Modelling in Management* 19(6):1907–1926, 2024. doi:10.1108/JM2-01-2024-0015.

[22] K. Subramanian, F. Hajamohideen, V. Viswan, N. Shaffi, and M. Mahmud. Exploring intervention techniques for Alzheimer's disease: Conventional methods and the role of AI in advancing care. *Artificial Intelligence and Applications* 2(2):59–77, 2024. doi:10.47852/bonview42022497.

[23] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018. doi:10.1109/CVPR.2018.00931.

[24] K. Tan, J. Wu, H. Zhou, Y. Wang, and J. Chen. Integrating advanced computer vision and AI algorithms for autonomous driving systems. *Journal of Theory and Practice of Engineering Science* 4(1):41–48, 2024. doi:10.53469/jtpes.2024.04(01).06. `https://centuryscipub.com/index.php/jtpes/article/view/427`.

[25] P. Tank. Social media post dataset. Kaggle Dataset, 2024. `https://www.kaggle.com/datasets/prishatank/post-generator-dataset/data`.

[26] F. Tiago. ecommerce-product-dataset. GitHub Repository, 2025. `https://github.com/octaprice/ecommerce-product-dataset`.

[27] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, et al. SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems* 36:8815–8827, 2023. doi:10.48550/arXiv.2305.02034.

[28] Y. Xue and J. Williams. Inducing shifts in attentional and preattentive visual processing through brief training on novel grammatical morphemes: An event-related potential study. *Language Learning* 74(S1):185–223, 2024. doi:10.1111/lang.12642.

[29] P. Zhang, J. Zheng, H. Lin, C. Liu, Z. Zhao, et al. Vehicle trajectory data mining for artificial intelligence and real-time traffic information extraction. *IEEE Transactions on Intelligent Transportation Systems* 24(11):13088–13098, 2023. doi:10.1109/TITS.2022.3178182.

[30] X. Zheng, X. Qiao, Y. Cao, and R. W. Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics* 38(4):133, 2019. doi:10.1145/3306346.3322971.

[31] X. Zhong, J. Tang, and A. Jimeno Yepes. PubLayNet: Largest dataset ever for document layout analysis. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1015–1022, 2019. doi:10.1109/ICDAR.2019.00166.