Automatic Extraction Algorithm For Visual Communication Design Elements Based On Deep Learning

Yinjie Zhao*

School of Art Education, Hubei Institute of Fine Arts, Wuhan, Hubei 430205, China

Corresponding author. E-mail: zhaoyinjie28@sina.com

Received: Mar. 25, 2026; Accepted: May. 27, 2026

This paper addresses the challenge that traditional visual communication design algorithms struggle to extract reliable features from complex design elements. To address this issue, an automatic extraction algorithm based on deep learning is proposed, improving the intelligence and efficiency of visual communication design. Furthermore, the YOLOv7 target recognition model is enhanced by integrating a Squeeze-and-Excitation (SE) attention mechanism, which improves its capability to detect small and complex visual elements. The improved model is trained and thoroughly evaluated using experimental data. This study confirms that deep learning-based automatic extraction methods can effectively support visual communication design and provide valuable insights for its intelligent development.

Keywords: deep learning; visual communication; design elements; automatic extraction

http://dx.doi.org/10.6180/jase.202612_35.003

1. Introduction

Visual communication design is the design of information transmission with the help of media tools. Nowadays, people are in the era of Internet development, and visual communication design takes the network as its activity space to carry out design activities in various fields to meet the needs of all kinds of people. As an active user group of the Internet, youth fun community has special requirements for visual communication design. This group and visual communication design interact with each other through the network. Moreover, as the main force of the network, they pin their emotions on the network, forming a unique youth interest culture. Visual communication design uses media and the internet to convey information [1]. Computer-aided design software is widely used for creating 3D models and 2D renderings in visual communication [2]. There are many natural restrictions for conventional approaches to extract design elements from visual communication designs. They can be summarized into three categories.

Visual Communication Design finds extensive application across various disciplines. In advertising, visual communication enhances appeal and strengthens brand-audience interaction. This paper enhances YOLOv7 using an SE attention mechanism to improve small target detection in visual elements, evaluation, and experimental validation. Classical visual communication algorithms rely on manual features, limiting robustness and generalization. Deep learning improves feature learning but still faces challenges in small object detection. This paper combines deep learning algorithms to improve the YOLOV7 target recognition method, adds SE attention mechanism to make the improved model more suitable for the detection and extraction of small target features in visual communication elements, trains the target detection model and evaluates the training results. In addition, this paper carries out the experiment of visual communication element feature target recognition.

2. Related works

(1) Visual information processing: Visual information is perceived from the environment and recognized through stored memory [3]. KMS research focuses on knowledge creation, sharing, and reuse but often lacks a complete framework [4].
(2) Integration of image and visual communication art: Images enhance visual communication by conveying information effectively [5]. To improve the robustness of visual models used in image-based communication systems, data augmentation techniques are commonly applied during preprocessing. Rotation improves orientation recognition; noise enhances robustness and generalization. Reference [6] pointed out that according to the investigation and practice, dynamic visual systems outperform traditional design in recognition and transmission.
(3) Variable visual identity design: Reference [7] explores variable brand identity design balancing identity and diversity in branding effectively. Reference [8] applies Gestalt principle, categorizing variable brand identity into four adaptive morphological design types. Unlike other attention-based object detection techniques, the IM-YOLOV7 algorithm implements a dual-branch framework that considers both the spatial features and semantic dependencies simultaneously. These studies offer theoretical basis but lack practical implementation details for variable brand identity design, causing fragmentation between theory and application.

3. Algorithm model

3.1. Feature recognition

Deep learning detection models include one-stage (YOLO, SSD) for speed but lower accuracy, and two-stage (R-CNN, Faster R-CNN) for higher accuracy using candidate regions. YOLOv7 is a one-stage detector predicting bounding boxes and class probabilities, improving accuracy and efficiency through enhanced feature aggregation compared to earlier versions.

Fig. 1 shows YOLOv7 with backbone for hierarchical feature extraction, neck for multi-scale feature fusion, and detection head for predicting bounding boxes, confidence, and class probabilities. YOLOv7 has average performance in small target detection, so this paper improves the YOLOv7 target detection model and adds SE attention mechanism. The SE attention module is integrated after each convolutional block in the backbone network. Squeeze uses global average pooling and fully connected layers to capture channel dependencies; excitation applies sigmoid weights for adaptive feature recalibration of input maps. In

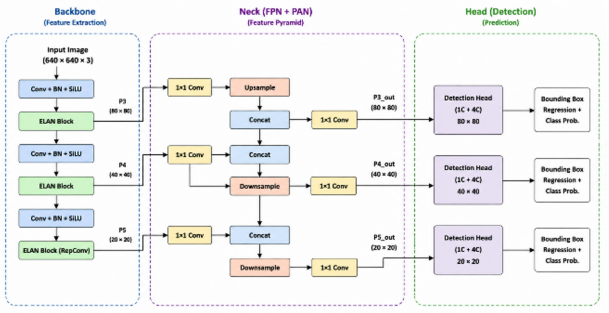

Fig. 1. Architecture of YOLOv7 showing backbone, neck, and detection head for multi-scale object detection

the squeeze stage, global average pooling summarizes each channel into a feature vector. The principle of SE is shown in Fig. 2. Attention mechanism can be intuitively explained as a way by which the network decides on how to give more importance It assigns higher weights to important channels and lower to irrelevant ones, improving feature discrimination and enhancing performance on complex, fine-grained visual communication datasets.

In addition to its computational function, the SE attention can also be considered as a feature recalibration method channel by channel

IM-YOLOv7 integrates SE-enhanced backbone and dual-branch design combining YOLOv7 detection with TCN-GRU-Self-Attention for parallel spatial and semantic feature extraction.The IM-YOLOv7 model proposed in this article adopts a hierarchical fusion architecture:

1. Bottom level feature extraction: Add SE attention modules after each convolutional layer in the YOLOv7 backbone network to enhance sensitivity to small targets;

2. High level semantic fusion: The TCN+GRU+Self Attention combination module is used as a parallel branch to parse the semantic associations of design elements. The TCN+GRU+Self Attention is designed as a sequential feature modeling framework. TCN captures long-range dependencies, GRU models sequences, and self-attention highlights important temporal features using weighted hidden states effectively.

The data augmentation includes rotation, flipping, Gaussian noise, and multi-sample enhancement to improve visual communication image diversity and robustness [9].

1. Principle of image rotation: The principle of image rotation is similar to the rotation process of a point in the rectangular coordinate system around the coordinate

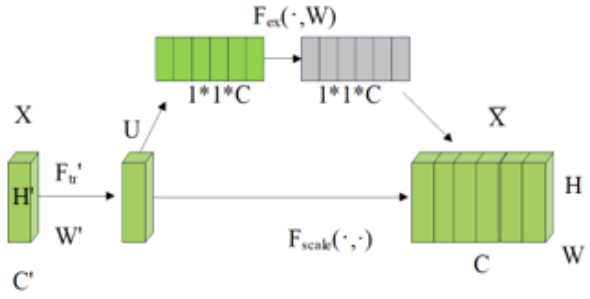

Fig. 2. SE attention mechanism illustrating channel-wise feature recalibration using squeeze and excitation operations

origin. As shown in Fig. 3, $P(x,y)$ in the rectangular coordinate system is rotated around the origin to $Q(x',y')$.

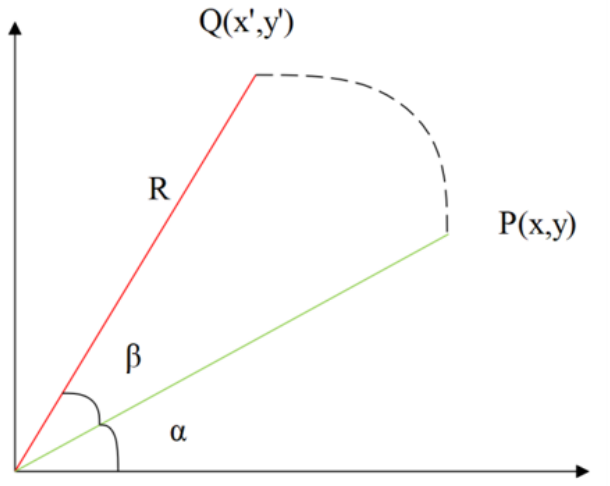

Fig. 3. Rotation process diagram

Since the distance to the coordinate origin does not change after rotation, the distance from P to the coordinate origin is assumed to be R. Then, there are:

$$R = \sqrt{x^2 + y^2} = \sqrt{x'^2 + y'^2} \tag{1}$$

$$\begin{cases} x' = R\cos(\alpha+\beta) = x\cos\beta - y\sin\beta \\ y' = R\sin(\alpha+\beta) = x\sin\beta + y\sin\beta \end{cases} \tag{2}$$

In digital image processing, images are stored in matrix form, so the above formula can be expressed as:

$$R = \begin{bmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

2. Adding Gaussian noise to pictures of visual communication works: The probability density function of Gaussian noise obeys a normal distribution, which has two parameters mean u and standard deviation σ, and the expression is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{x-u}{2\sigma^2}\right) \tag{4}$$

YOLO dataset uses Images and Annotation folders; Labeling annotates industrial design works focusing on symbols, color zoning, and layout structure [10]. The proposed automatic extraction algorithm is organized into a sequential processing pipeline to provide a clear understanding of the overall framework. Images are preprocessed, features extracted using improved YOLOv7 with SE attention, detections predicted, and post-processing applied using thresholding and non-maximum suppression.

The improved detection architecture is integrated with the automatic extraction process through a modular decomposition strategy. Each module improves feature refinement; SE-enhanced YOLOv7 ensures robustness for small targets.

After completing the research and improvement of the target detection model and the production of the data set of visual communication works. The training process is shown in

3.2. Stereo matching

Stereo matching, also called disparity estimation, finds corresponding pixels in left and right images to generate a disparity map, representing depth differences in a 3D scene. When considering the extraction of elements of visual communication design, stereo matching enhances feature representation by exploiting spatial and structural relationships. Depth and disparity information help distinguish overlapping visual elements and improve separation of similar objects. Stereo matching uses disparity and camera calibration to estimate depth, with traditional methods involving cost computation, aggregation, and optimization.

1. Calculation of matching cost: Matching cost determines correspondence between left and right image pixels, providing a quantitative measure for stereo matching accuracy.

2. Cost aggregation: Cost aggregation improves stereo matching accuracy by refining pixel costs using neighboring information, mainly in local and semi-global methods.

These parts of the existing stereo matching algorithms are large together. Commonly used stereo matching algorithms include matching based on AD, SAD, Census, etc.

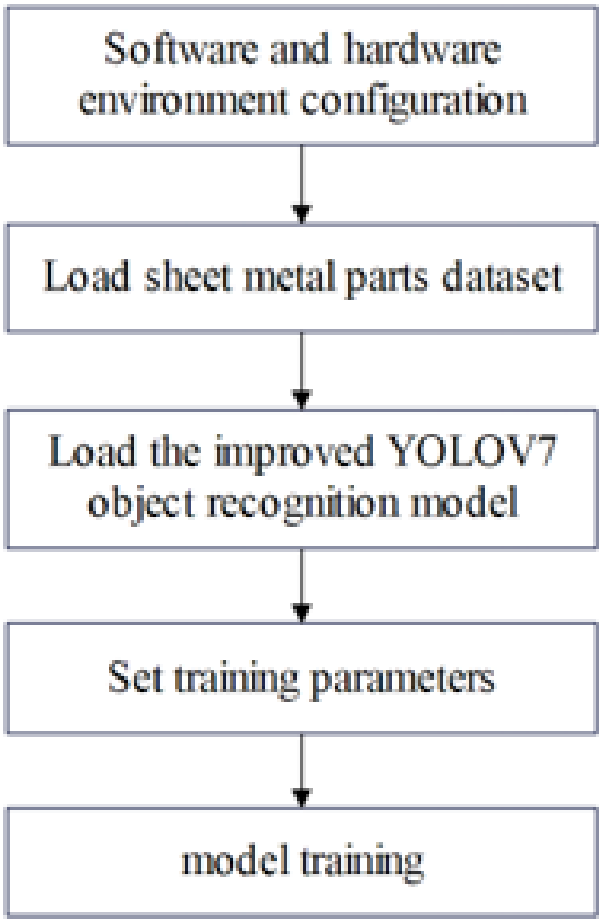

Fig. 4. Training process

[11]. (1) AD: The simplest matching cost method compares grayscale values between left and right images by fixing a pixel, traversing corresponding pixels, and repeating the process.

$$C_{AD}(p,q) = |I_L(p) + I_R(q)| \tag{5}$$

Among them, p and q are the coordinates of the pixel points, and $I_L(p)$ and $I_R(q)$ are the grayscale values. If the image is a color image, there is the following calculation formula:

$$C_{AD}(p,q) = \frac{1}{3}\sum_{i=R,G,B}\left|I_i^L(p) + I_i^R(q)\right| \tag{6}$$

(2) SAD algorithm: SAD is also a commonly used matching cost calculation method. The basic flow of SAD algorithm is as follows: 1) The algorithm inputs two rectified left view L and right view R. 2) The algorithm scans the left view L and selects a position to build a sliding window, similar to the convolution kernel in convolution processing. 3) The algorithm covers L with this window and extracts the pixel values in the area. 4) The algorithm extracts the pixel values of the corresponding area of R again.

$$C_{SAD}(p,q) \sum_{m\in N_p, n\in N_q} |I_L(m) + I_R(n)| \tag{7}$$

(3) Census transform method: Census transform method is another widely used stereo matching cost calculation method. The transformation process is shown in Fig. 5. The transformation formula is:

$$T(q) = \bigotimes_{m\in N_p} \xi(I(p), I(q)) \tag{8}$$

Among them, p is the center pixel of the window, q is the pixels other than the center pixel of the window, and N_p represents the neighborhood of the center pixel p [12].

The matching cost calculation method based on Census transform is to calculate the Hamming distance of the Census transform values of the two pixels corresponding to the left and right images, that is, the matching cost is [13]:

$$C_{\text{Cencus}}(p,q) = \text{Ham min}\, g(T(p), T(q)) \tag{9}$$

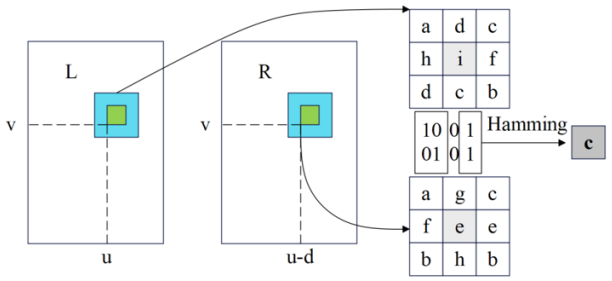

Fig. 5. Census transformation process

4. Systems and tests

4.1. Model construction

A Java-based B/S architecture system with MySQL storage is developed for style salience image classification, enabling efficient style similarity retrieval and prediction with accurate processing. The suggested system developed bearing scalability and real-time applicability in mind. The YOLOv7 one-stage model with attention mechanism enables efficient analysis of large visual communication datasets, supports parallel computation, reduces cost, and maintains robust performance across varying data distributions. The system follows MVC architecture: view handles UI display, controller manages HTTP requests and routing, and model performs data operations including efficient reading and storage. The schematic diagram of its system technical architecture is shown in Fig. 6.

Fig. 6. Schematic diagram of system technical architecture

The system uses modular design based on requirements analysis to improve implementation efficiency and better meet user needs. The overall functional architecture diagram of the system is shown in Fig. 7.

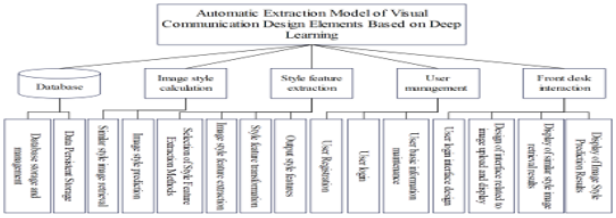

Fig. 7. Overall functional architecture diagram of the system

The high system performance, clear definitions of the development and operating environment requirements are necessary, as they are critical for system operation and efficiency. The system development environment configuration of this paper is as follows:

Operating system: Windows11

Database: MySQL

Programming languages: Java, Html5, CSS, JavaScript

Frameworks: Flask, jQuery, Bootstrap

Development tool: Intelli JIDEA

Assistive tools: Navicat Premium15.

Model efficiency evaluated via inference time, runtime, and hardware; YOLOv7-SE adds minor overhead while maintaining real-time performance and scalability.

The system is designed for visual communication design element analysis, focusing on symbols, layout structures, and color regions. The system analyzes visual design elements using modular architecture for efficient processing, deployment, and visualization across diverse scenarios.

The dataset includes annotated visual communication elements from public and custom sources, with symbols and graphics, balanced via augmentation and YOLO-format normalized bounding box annotations.

4.2. Experimental results

IM-YOLOv7 uses 1e-3 learning rate with scheduled decay, early stopping, varied hidden layers, three dataset runs, and averaged results in Table 1.

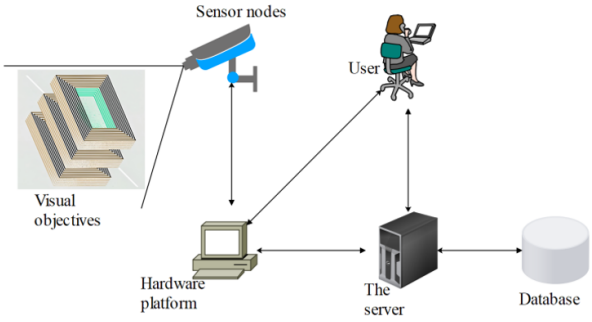

Fig. 8. Experimental design ideas

The Cityscapes public dataset is used to compare and evaluate the effects of different models under multi-modal feature data. The experimental results are shown in Table 2. The ablation study validates IM-YOLOv7 in two stages: Stage 1 evaluates TCN, GRU, and Self-Attention improvements.

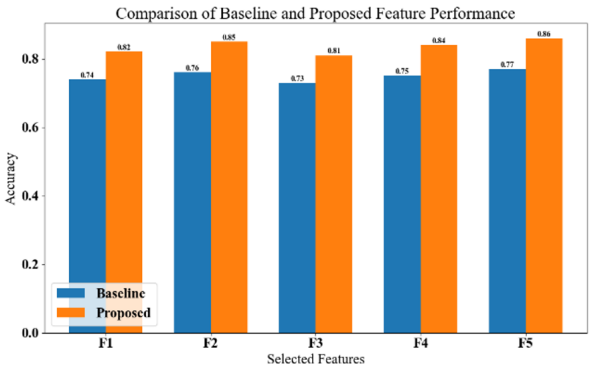

Fig. 9. Comparison of feature-level performance between the baseline model and the proposed IM-YOLOv7 model

Fig. 9 presents a visual comparison between the baseline model and the proposed IM-YOLOv7 across multiple feature sets.

Table 3 shows IM-YOLOv7 outperforms other models in F1 score, precision, recall, and accuracy, proving its effectiveness and robustness in visual communication detection.

Table 4 shows that progressively adding TCN, GRU, and Self-Attention modules consistently improves performance across all metrics, highlighting their contribution to better feature representation.

The contribution of individual feature components, a visual comparison of ablation configurations is presented in Fig. 10.

IM-YOLOv7 is tested via MATLAB simulation; fifteen groups evaluate user experience and feature recognition, with results shown in Table 5.

Table 1. Performance parameters of the model under different hidden layers

Sequence	Model	F1	Specificity	Precision	Recall	Accuracy
1	CNN	0.683	0.714	0.663	0.724	0.704
2	GRU	0.663	0.693	0.704	0.642	0.673
3	LSTM	0.785	0.806	0.765	0.806	0.795
4	TCN (all)	0.775	0.795	0.724	0.846	0.785
5	TCN (sentence)	0.806	0.826	0.755	0.867	0.816
6	CNN-Attention	0.744	0.775	0.693	0.795	0.765
7	GRU-Attention	0.836	0.867	0.795	0.877	0.857
8	BiAttention-GRU	0.918	0.928	0.928	0.908	0.908
9	IM-YOLOv7	0.938	0.959	0.959	0.918	0.928

Table 2. F1 value, specificity, precision, recall and accuracy indicators of multimodal data

Sequence	Layer	F1	Specificity	Precision	Recall	Accuracy
1	3	0.530	0.612	0.581	0.489	0.561
2	4	0.795	0.693	0.724	0.877	0.755
3	5	0.938	0.959	0.959	0.918	0.928
4	6	0.918	0.928	0.928	0.908	0.897
5	7	0.918	0.928	0.928	0.908	0.908
6	8	0.908	0.918	0.918	0.897	0.897

Table 3. Performance comparison with state-of-the-art object detection models

Model	F1 Score	Precision	Recall	Accuracy
YOLOv7 [10]	0.918	0.928	0.908	0.908
YOLOv8 [5]	0.925	0.935	0.910	0.915
DETR [7]	0.910	0.920	0.900	0.905
IM-YOLOv7	0.938	0.959	0.918	0.928

Table 4. Comparison of time performance and performance with different superpixel methods

Sequence	Model	F1	Specificity	Precision	Recall	Accuracy
1	TCN (all)	0.775	0.795	0.724	0.846	0.785
2	TCN (sentence)	0.806	0.826	0.755	0.867	0.816
3	TCN (sentence)+GRU	0.836	0.857	0.806	0.877	0.846
4	TCN (sentence)+GRU+Self-Attention	0.908	0.928	0.918	0.908	0.897
5	IM-YOLOv7	0.938	0.959	0.959	0.918	0.928

Table 5. User experience evaluation results

Number	User experience	Number	User experience	Number	User experience
1	0.925	6	0.878	11	0.926
2	0.900	7	0.906	12	0.884
3	0.916	8	0.891	13	0.881
4	0.929	9	0.890	14	0.882
5	0.874	10	0.883	15	0.899

Table 6. The evaluation process, mean Average Precision (mAP) and Intersection over Union (IoU) metrics

Model	mAP@0.5	mAP@0.5:0.95	IoU
YOLOv7	0.872	0.621	0.68
IM-YOLOv7	0.935	0.701	0.78

4.3. Analysis and discussion

IM-YOLOv7 outperforms CNN, GRU, LSTM, TCN in accuracy, precision, recall, and F1 score. Despite the promising performance of the proposed IM-YOLOv7 model, several limitations should be acknowledged. The study has limitations including dataset bias from general-purpose sources,

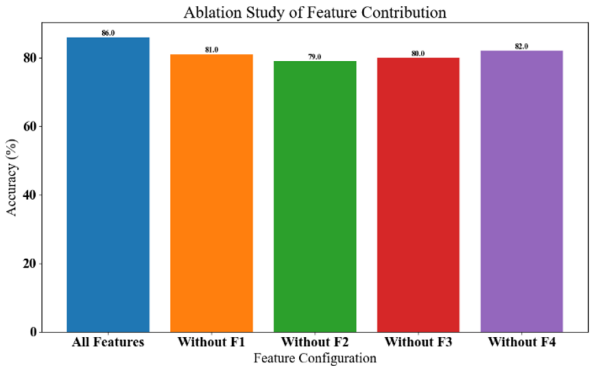

Fig. 10. Ablation study illustrating the contribution of individual feature components to overall model accuracy

restricted evaluation on industrial design elements, and sensitivity.

Table 3 shows IM-YOLOv7 improves performance stepwise with GRU, Self-Attention, and YOLOv7, enhancing accuracy, precision, recall, and F1 score.

Fig. 11. Accuracy improvement curve showing convergence behavior of the proposed IM-YOLOv7 model during training

The training convergence behavior of the proposed model is illustrated in Fig. 11. The performance trend demonstrates a steady improvement in accuracy over training iterations. IM-YOLOv7 improves feature representation through SE block enhancing small object features, TCN capturing structural relationships, and GRU modeling contextual dependencies, collectively refining sequential extraction and strengthening overall detection performance.

5. Conclusion

This paper proposed an automatic extraction algorithm for visual communication design elements based on an improved deep learning framework. The YOLOv7 model was enhanced by integrating the SE attention mechanism to improve the detection ability of small and complex visual elements. In addition, a hybrid feature extraction structure combining TCN, GRU, and Self-Attention modules was introduced to strengthen semantic and contextual feature learning. The proposed IM-YOLOv7 model effectively improved feature representation capability and detection robustness in complex visual communication scenarios. Experimental results showed that the proposed model achieved better F1-score, precision, recall, accuracy, and mAP values compared with existing baseline models such as CNN, LSTM, GRU, YOLOv7, YOLOv8, and DETR. The ablation experiments further confirmed that each added module contributed positively to the overall system performance and feature discrimination ability.

6. Declarations

7. Funding

There is no specific funding to support this research.

8. Conflict of interest

The authors declare that they have no conflicts of interest regarding this work.

9. Data availability

The datasets generated during and or analyzed during the current study are available from the corresponding author on reasonable request.

10. Code availability

Not applicable.

11. Ethical approval

This study does not involve any human participants or animals

12. Consent to participate

Not applicable. The study does not involve human participants, and no consent to participate.

13. Consent to publish

Not Applicable. This study does not include human participants or identifiable images.

14. Author contributions

All Author contributed to the design and methodology of this study, the assessment of the outcomes, and the writing of the manuscript.

References

[1] L. Lu, (2020) *"Design of visual communication based on deep learning approaches"* **Soft Computing *24***: 7861–7872. DOI: 10.1007/s00500-019-03954-z.

[2] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, (2022) *"Toward semantic communications: Deep learning-based image semantic coding"* **IEEE Journal on Selected Areas in Communications *41***: 55–71. DOI: 10.1109/JSAC.2022.3221999.

[3] G. Zhang, B. Liu, T. Zhu, A. Zhou, and W. Zhou, (2022) *"Visual privacy attacks and defenses in deep learning: a survey"* **Artificial Intelligence Review *55***: 4347–4401. DOI: 10.1007/s10462-021-10123-y.

[4] D. S. Hidayat, D. I. Sensuse, D. Elisabeth, and L. M. Hasani, (2025) *"Conceptual model of knowledge management system for scholarly publication cycle in academic institution"* **VINE Journal of Information and Knowledge Management Systems *55***(1): 187–222. DOI: 10.1108/VJIKMS-08-2021-0163.

[5] D. Ivanko, D. Ryumin, and A. Karpov, (2023) *"A review of recent advances on deep learning methods for audio-visual speech recognition"* **Mathematics *11***: 2665. DOI: 10.3390/math11122665.

[6] X. Yang, X. Li, Y. Guan, J. Song, and R. Wang, (2020) *"Overfitting reduction of pose estimation for deep learning visual odometry"* **China Communications *17***: 196–210. DOI: 10.23919/JCC.2020.06.016.

[7] W. Zhou, J. Lei, Q. Jiang, L. Yu, and T. Luo, (2020) *"Blind binocular visual quality predictor using deep fusion network"* **IEEE Transactions on Computational Imaging *6***: 883–893. DOI: 10.1109/TCI.2020.2993640.

[8] T. J. Thomson, D. Angus, P. Dootson, E. Hurcombe, and A. Smith, (2022) *"Visual mis/disinformation in journalism and public communications: current verification practices, challenges, and future opportunities"* **Journalism Practice *16***: 938–962. DOI: 10.1080/17512786.2020.1832139.

[9] X. Liu and R. Yao, (2023) *"Design of visual communication teaching system based on artificial intelligence and CAD technology"* **Computer-Aided Design and Applications *20***: 90–101. DOI: 10.14733/cadaps.2023.S10.90-101.

[10] Z. Nie, Y. Yu, and Y. Bao, (2023) *"Application of human–computer interaction system based on machine learning algorithm in artistic visual communication"* **Soft Computing *27***: 10199–10211. DOI: 10.1007/s00500-026-11364-1.

[11] L. Fernández, (2020) *"The emotional politics of images: moral shock, explicit violence and strategic visual communication in the animal liberation movement"* **Journal of Critical Animal Studies *17***: 53–80.

[12] L. Sun, P. Wang, P. Liu, and Z. Nie, (2023) *"Image processing method of a visual communication system based on convolutional neural network"* **International Journal of Semantic Web and Information Systems *19***: 1–19. DOI: 10.4018/IJSWIS.330022.

[13] Y. Wang, (2022) *"Illustration art based on visual communication in digital context"* **Mobile Information Systems *2022***: 7364003. DOI: 10.1155/2022/7364003.