ISSN: 3104-5235

INNO PRESS

July 2025

Journal of Emerging Applied Artificial Intelligence

Volume 1 / Issue 4

Issue 4 – Foundations of Emerging Applied Artificial Intelligence

The Journal of Emerging Applied AI (JEAAI) is pleased to present its inaugural issue, establishing a dedicated forum for high-quality, peer-reviewed scholarship at the intersection of artificial intelligence theory and real-world application. This first issue reflects the journal's foundational mission: to advance and disseminate research that demonstrates the transformative potential of AI technologies across sectors and disciplines.

This opening volume features contributions that exemplify the journal's emphasis on rigorously developed, practically deployed AI systems. The selected articles cover a spectrum of domains—including healthcare, robotics, transportation, education, and sustainability—demonstrating the breadth of AI's impact when translated from conceptual innovation to applied implementation.

With a commitment to methodological soundness, interdisciplinary relevance, and societal benefit, JEAAI aims to become a leading platform for scholars, practitioners, and innovators who are engaged in solving real-world problems through intelligent systems. The journal's scope encompasses original research, technical reports, case studies, and critical perspectives, all grounded in applicability and reproducibility.

We invite the academic and professional community to engage with JEAAI as contributors, reviewers, and readers, and to join us in shaping a future where applied artificial intelligence drives meaningful and responsible progress.

Editor-in-Chief

Section Editors

Dr. Meng Liu
Research Scientist, NVIDIA

Dr. Liu's research interests include graph neural networks, clustering, and multimodal learning. He has published over 20 papers in leading venues such as *Advanced Science*, *IEEE TPAMI*, *IEEE TKDE*, *CVPR*, *ICML*, and *ICLR*. His work includes an ESI Hot Paper and a Highly Cited Paper, with over 1,000 citations. He has received several awards, including Best Paper at the 2024 China Computational Power Conference and a DAAD AInet Fellowship.

Dr. Zhongbin Luo
Professor-level Senior Engineer, China Merchants Chongqing Communications Research & Design Institute. Master's Supervisor, Chongqing Jiaotong University & Shijiazhuang Tiedao University

Dr. Luo's research focuses on intelligent transportation, traffic safety, and vehicle–road collaboration. He has led over ten national and provincial research projects, holds 11 invention patents, and serves as an expert reviewer for journals such as *IEEE Access* and *PLOS ONE*.

Dr. Ruichen Xu
Postdoctoral Fellow, Department of Civil & Environmental Engineering,University of Missouri, Columbia, USA

Dr. Xu's research interests include hydrological ecology, AI-based flood forecasting, and sediment–pollutant dynamics. He has led or contributed to more than ten projects in China and the U.S. and has published over 20 peer-reviewed papers. He holds patents in environmental monitoring and serves as a reviewer for journals like *Journal of Hydrology* and *Ecological Indicators*.

A/Prof. Jinghao Yang

Assistant Professor, Electrical and Computer Engineering, The University of Texas Rio Grande Valley, USA

Dr. Yang has taught in the U.S. and specializes in applying machine learning to intelligent manufacturing systems. His research bridges intelligent sensing, control, and adaptive design with industrial applications, contributing to smart production technologies and data-driven innovation.

Luxin Zhang
PhD Candidate, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, New Zealand

Luxin Zhang is currently pursuing her PhD in Artificial Intelligence. Her research focuses on machine learning algorithms and their applications in intelligent systems. As Managing Editor, she is responsible for manuscript assignment, editorial coordination, and issue scheduling. Based in New Zealand, she serves as a central figure in the journal's daily operations.

Yihan Zhao
PhD Candidate, University of Auckland, New Zealand

Yihan Zhao holds a Master’s degree from Peking University and is currently a PhD candidate at the University of Auckland. Her research explores the intersection of communication, culture, and technology, with a focus on how algorithms reshape cultural expression and the subjectivity of marginalized communities. She previously served as an Assistant Research Fellow at the Development Research Centre of the State Council in China, contributing to national research projects. She has curated and coordinated panels for the China Development Forum, facilitating high-level dialogue on AI, sustainability, and governance.

Shen (Jason) Zhan
Graduate Researcher, University of Melbourne, Australia

Jason Zhan holds an Honours degree in Civil and Environmental Engineering from the University of Auckland and is currently a PhD researcher in the Teaching & Learning Lab at the University of Melbourne. He combines industry and academic experience, with a background in structural engineering and teaching. His research focuses on employability assessment and curriculum design in engineering education, with growing interest in the role of AI in authentic assessment and personalized learning.

Contents

1. BGC-YOLO: A Feature Fusion-Based Algorithm for Traffic Sign Detection 1

2. Hybrid Causal-Predictive Framework for Data Asset Valuation and Regulatory-Integrated Financial Reporting in Manufacturing Enterprises 10

3. AI-Powered Two-Phase Method for Microscopic Periodic Railway Operation Diagrams 17

4. Research on the Application of Artificial Intelligence Product Design in Human Emotions: A Case Study of Chinese Women 33

5. Innovations and Frontiers of Diffusion Models in Natural Language Processing: A Review .. 46

BGC-YOLO: A Feature Fusion-Based Algorithm for Traffic Sign Detection

Shuo Cui[1,2], YingZhao Xue[1,2], and ZeKai Liu[1,2]
[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong, China
[2] Shanxi Provincial Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Jinzhong, China

Abstract—With the development of intelligent transportati on systems, the automatic detection of traffic signs has become a key task in assisted driving and unmanned driving perceptio n systems. In view of the problem that traffic signs are small in scale in images and their accuracy is affected by complex envi ronments, this paper constructs a BGC-YOLO target detection algorithm based on YOLOv11. First, by introducing the bidire ctional feature fusion structure BiFPN, the interactive expressi on of multi-scale features is enhanced. Secondly, the global-lo cal spatial attention mechanism GLSA is combined to improve the model's perception of detail information and contextual se mantics. Finally, the content-aware upsampling module CARA FE is used to optimize the feature reconstruction process and e ffectively retain the key information of small targets. The expe rimental results on the CCTSDB2021 traffic sign dataset show that the improved model achieves a good balance between acc uracy and efficiency, with an increase of 1.4% in mAP@0.5 co mpared to the original model, and maintains a low computatio nal overhead, which is practical.

Index Terms—Traffic sign detection, YOLOv11, feature fusion, object detection

I. INTRODUCTION

Traffic signs, as the core carrier of road information, play a vital role in ensuring the safe driving of vehicles. They are also an indispensable key link in realizing autonomous driving technology. Due to the wide variety of traffic signs, they often appear in complex and changing background environments. In addition, the system requires real-time detection results, making automatic detection and recognition of traffic signs a very challenging task.

Early traffic sign detection methods mainly rely on artificially designed features such as color and shape to achieve classification and recognition. For example, Bahlmann[1] proposed a method that uses color, shape and motion information for traffic sign detection; Li H[2] combined color segmentation and robust shape matching with a new method and used support vector machines for classification. Although these traditional methods have achieved results to a certain extent, they generally rely on specific manual feature design for different traffic signs and are easily affected by environmental noise, resulting in poor robustness. To overcome these limitations, researchers began to introduce deep learning models into traffic sign detection tasks[3]. Compared with traditional methods, deep learning-based models have become the mainstream technical path in the field of traffic sign detection in recent years due to their higher recognition accuracy and stronger anti-interference ability.

At present, the object detection methods based on deep learning are mainly divided into two categories: two-stage detection algorithms represented by the R-CNN[4] series and single-stage detection algorithms represented by YOLO[5]. The two-stage method usually generates candidate regions first and then performs classification and regression. Although it performs well in detection accuracy, it is relatively slow due to the complex process. In contrast, the single-stage algorithm omits the step of generating candidate regions, which can achieve faster detection speed and is suitable for real-time applications. However, it still has certain shortcomings in detection accuracy, especially in processing small objects.

In order to further break the limitations of traditional convolutional architecture in modeling long-distance dependencies and object relationships, the DETR model was proposed[6], which introduced the Transformer architecture to build a new end-to-end object detection framework. DETR transforms the object detection task into a set prediction problem, no longer relying on candidate region generation or non-maximum suppression, and realizes the modeling of global image information through the self-attention mechanism. This method shows significant advantages in modeling object relationships and complex semantic contexts, and is particularly suitable for optimizing object position and category prediction in dense scenes. However, DETR still has shortcomings in convergence speed and small target detection, which has prompted the proposal of a series of improved variants to balance detection accuracy and training efficiency.

Aiming at the problem of missed detection of small targets when the span of traffic signs is large, as well as the problem of

This research was funded by the Shanxi Provincial Science and Technology Strategic Research Special Key Project (Project No.: 202304031401011) and the Shanxi Provincial Basic Research Plan (Free Exploration) Project (Project No.: 202403021222276). The corresponding author is Shuo Cui (email: cs811767@163.com). The author YingZhao Xue (email: 18366902381@163.com) and ZeKai Liu(email: 15536368230@163.com) are from the School of Computer Science and Technology, Taiyuan Normal University.

false detection in complex environments, this paper takes the YOLOv11n model as the basic architecture from the perspective of improving detection accuracy and robustness, comprehensively considers the detection speed and deployment efficiency of the model, and proposes a BGC-YOLO traffic sign detection model. The work of this paper is as follows:

1) In order to enhance the multi-scale feature fusion capability, the BiPFN feature fusion network is introduced. Through richer bidirectional paths and cross-layer connections, the feature expression capability of targets of different scales, especially small target traffic signs, is improved.
2) The GLSA attention mechanism is introduced in the Neck part to enhance the information selectivity of the model in the feature fusion process. GLSA pays attention to local details and global context at the same time. By weighted selection of semantic features at different levels, it effectively improves the model's perception of the edge and shape details of traffic signs and improves the accuracy of target recognition under complex background interference.
3) The lightweight and efficient CARAFE module is used to replace the original nearest neighbor interpolation method. CARAFE achieves more accurate high-resolution feature reconstruction through content-aware reconstruction mechanism, effectively preserving the detailed information of small target traffic signs.

II. Related Works

A. R-CNN Series Object Detectors

The R-CNN family has had a significant impact on the evolution of deep learning-based object detection frameworks. The original R-CNN framework was proposed by Girshick et al. in 2014. It proposed a two-stage detection process: using selective search to generate region proposals, each region is independently passed through a CNN to extract features, and then classified and bounding box regression is performed. Although R-CNN shows high detection accuracy, its computational efficiency is low due to the redundant forward propagation of thousands of regions, which poses a challenge for real-time applications.

To overcome these limitations, Fast R-CNN[7] was born, which processes the entire image only once through the convolutional backbone network. Then, region of interest (RoI) pooling is used to map region proposals to feature maps, which significantly improves speed and reduces memory usage. Faster R-CNN[8] further improves on this by introducing a region proposal network to generate region proposals directly from shared convolutional features, thereby building an end-to-end trainable detection system with state-of-the-art accuracy and higher efficiency.

Later advances, such as Mask R-CNN[9], extended Faster R-CNN by adding parallel branches, demonstrating the adaptability of the R-CNN family to more complex visual tasks. Other variants, such as Cascade R-CNN[10], Libra R-CNN[11], and R-FCN[12], further optimized multi-stage training, balanced feature representation, and fully convolutional reasoning to improve detection performance (both precision and recall).

Overall, the R-CNN family represents the foundational paradigm for two-stage object detection, known for its strong accuracy and scalability. However, computational complexity and inference speed remain limiting factors for real-time and resource-constrained applications, such as autonomous driving or embedded traffic sign detection systems.

B. YOLO Series Object Detectors

The YOLO family represents one of the most influential research directions in the field of real-time object detection. Unlike two-stage detectors such as R-CNN, the YOLO family adopts a single-stage end-to-end framework that can directly predict the object category and bounding box in the entire image in a single network transmission. This unified architecture significantly improves the inference speed, making YOLO particularly suitable for real-time applications such as autonomous driving and video surveillance.

The first version of YOLO, YOLOv1[13], was proposed by Redmon et al. It defines object detection as a regression problem, dividing the input image into a fixed grid and predicting the bounding box and category probability based on each grid cell. Although YOLOv1 exhibits impressive speed, it has poor localization accuracy and has difficulty detecting small or clustered objects.

To overcome these limitations, YOLOv2[14] introduced anchor boxes, batch normalization, and a new backbone network, which significantly improved accuracy without sacrificing speed. YOLOv3[15] further improved this performance by using multi-scale prediction and a deeper backbone network (Darknet-53), achieving a good balance between detection performance and inference speed for objects of different sizes.

YOLOv4[16], developed by Bochkovskiy et al., strikes a balance between accuracy and deployability, incorporating a variety of modern training strategies such as cross-stage partial connections (CSP), Mish activation function, and CIoU loss function. It achieves state-of-the-art performance on the COCO benchmark and has good generalization ability and efficiency.

The advent of YOLOv5[17] marked the transition of the model to a PyTorch-based implementation, which makes the model more widely adopted and easier to customize. YOLOv5 introduced a series of lightweight models and emphasized the feasibility of practical deployment by focusing on speed, scalability, and compatibility with various platforms.

YOLOv6[18], YOLOv7[19], and YOLOv8[20] further pushed the boundaries. YOLOv6 improves the neck and head design for industrial applications. YOLOv7 proposes an extended E-ELAN module and auxiliary head to improve detection accuracy and convergence speed. YOLOv8, developed by Ultralytics, focuses on unified tasks (detection, segmentation, pose estimation), adopts anchor-free detection head and decoupled head design to achieve better performance on different data sets.

The evolution from YOLOv9 to YOLOv13 continues to push lightweight object detection technology to break through the

limits. YOLOv9[21] optimizes feature extraction and gradient flow through programmable gradient information (PGI) and general efficient layer aggregation network (GELAN), achieving millisecond-level response on edge devices; YOLOv10[22] eliminates NMS dependence with an end-to-end architecture and combines spatial channel decoupling and downsampling technology to achieve a new benchmark for real-time detection on edge devices; YOLOv11[23] introduces a dynamic detection head to significantly improve the ability to parse complex scenes while maintaining its lightweight; YOLOv12[24] achieves global semantic modeling with extremely low computational cost by relying on regional attention (A2) and Flash Attention mechanisms; the latest YOLOv13[25] breaks the constraints of traditional architecture with HyperACE and FullPAD, and demonstrates excellent energy efficiency in scenarios such as drone inspection and smart wearables, promoting the development of target detection technology towards a more edge and real-time direction. In summary, the YOLO family has evolved from a basic real-time detector to a highly optimized family of robust models that achieve an excellent balance between speed and accuracy. Due to their simplicity, scalability, and computational efficiency, these detectors have become the cornerstone of many object detection systems.

C. DETR Series Object Detectors

The introduction of the Transformer architecture has brought a new research paradigm to the field of object detection. The most representative work is the DETR model proposed by Carion. DETR first applied the Transformer encoder-decoder structure to the object detection task, innovatively transformed the detection problem into a set prediction task, and omitted the candidate region generation and non-maximum suppression (NMS) modules commonly used in traditional methods. It uses the self-attention mechanism to model the global features of the image, and has good end-to-end trainability and structural simplicity. However, DETR has problems such as slow convergence speed and insufficient detection ability for small targets, which limits its wide deployment in practical applications.

To address these shortcomings, researchers have made many improvements to the DETR model and formed multiple variants, forming the DETR series of detectors. Among them, Deformable DETR[26] introduces a sparse multi-scale deformable attention mechanism, which enables the model to focus on local key positions, effectively improving the convergence speed and small target detection performance. Conditional DETR[27] uses a content-based dynamic query vector to enhance the model's adaptability to target semantics. DAB-DETR[28] introduces the idea of dynamic anchor boxes in the target query mechanism and improves positioning accuracy by iteratively optimizing the reference box position. Furthermore, DN-DETR[29] adopts a noise learning strategy to improve training stability and matching efficiency by introducing positive and negative sample noise. DINO[30], which combines the above optimization strategies, achieves a dual improvement in detection accuracy and training efficiency, and achieves excellent performance on multiple benchmark datasets. In addition, H-DETR[31] further enhances the local perception ability of the model by introducing the fusion of convolutional features and Transformer features, which is particularly suitable for dense small target scenes.

Overall, the DETR series of methods gradually make up for the limitations of the original model by introducing multi-scale features, dynamic query and auxiliary training mechanisms while maintaining the simplicity of structure and global modeling capabilities. It has become one of the key directions of Transformer architecture research in the field of target detection, and has shown broad application prospects in tasks such as autonomous driving, remote sensing image analysis, and traffic sign detection.

III. METHODS

A. BGC-YOLO Overview

In order to effectively improve the model's ability to detect multi-scale small targets in complex traffic environments, this paper proposes an improved detection model based on YOLOv11, BGC-YOLO. Its overall method introduces structural optimization and module integration to enhance performance from three dimensions: feature fusion, attention mechanism, and upsampling strategy. BGC-YOLO aims to solve problems such as complex background interference, insufficient expression of multi-scale features, and low accuracy in small target recognition.

First, in the feature fusion network, BGC-YOLO uses BiPFN (Bidirectional Pyramid Feature Network) to replace the PANet structure in the original YOLOv11. BiPFN introduces a bidirectional feature transfer mechanism to establish a more sufficient information flow between high-level semantic features and low-level detail features, thereby enhancing the model's ability to detect targets of different scales, especially small-sized traffic signs.

Secondly, the GLSA (Global-Local Selective Attention) attention mechanism is introduced in the Neck part to improve the model's ability to select information during feature fusion. Compared with the traditional attention module, GLSA integrates global context and local detail features, which can effectively highlight the key areas related to the target under complex backgrounds, and improve the discriminability and robustness of feature expression.

In addition, the model uses the CARAFE module to replace the original nearest neighbor interpolation or deconvolution method in the upsampling stage. CARAFE adaptively reconstructs spatial features through the content-aware dynamic convolution kernel generation mechanism, effectively retains the image structure details, and improves the restoration quality of small targets in high-resolution feature maps. This not only improves the model's perception of fine-grained targets, but also alleviates the common information loss problem in the upsampling process to a certain extent, which helps to improve the overall detection performance.

In summary, BGC-YOLO integrates three modules, BiPFN, GLSA and CARAFE, on the basis of the YOLOv11 framework,

and significantly enhances the model's robustness in complex backgrounds and the detection accuracy of small targets while maintaining the original detection speed advantage. The organic integration of the three in structure enables the model to complement each other in multi-scale feature modeling, contextual information attention, and detail feature reconstruction, building a more efficient, lightweight and expressive detection framework.

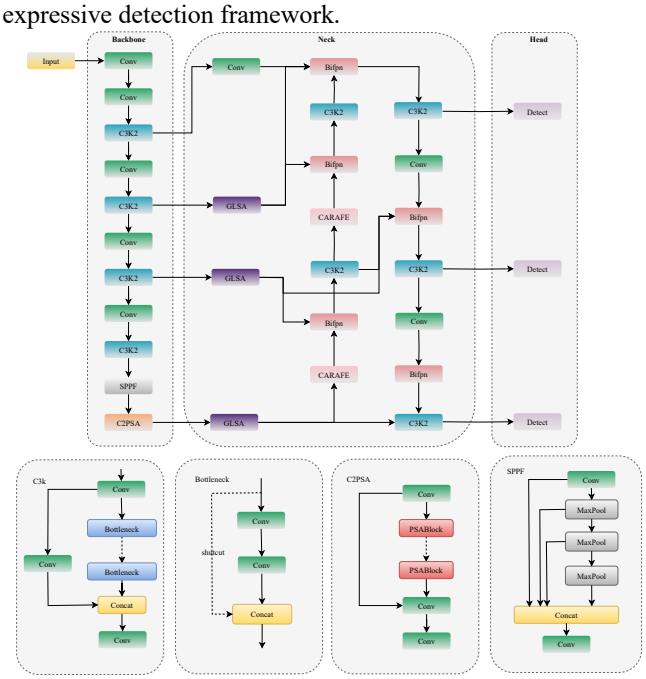


Fig. 1. BGC-YOLO structure

B. BiPFN

Although PANet can enhance the semantic transmission capability of features at different levels, it still suffers from problems such as insufficient feature flow and incomplete semantic fusion in small target detection scenarios. To address the problems of PANet in traffic sign detection tasks, such as low efficiency in small target feature transmission, insufficient fusion of upper and lower layer information, and lack of dynamic feature control capabilities, BiFPN[32] was used to replace the original PANet structure.

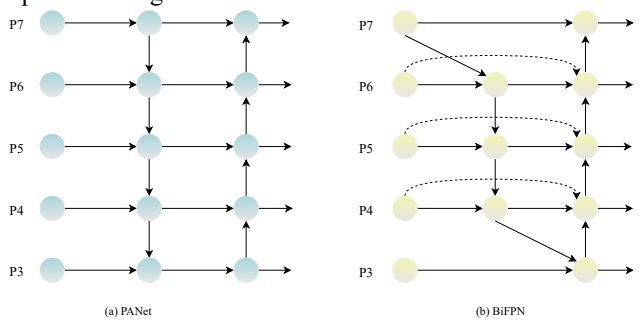


Fig. 2. BiFPN structure

As shown in Figure 2, BiFPN achieves bidirectional information flow between features at different levels by constructing bidirectional paths from top to bottom and from bottom to top, effectively enhancing the ability to integrate high-level semantics with low-level details. At the same time, the introduced learnable weighting mechanism allows the model to dynamically assign the importance of different feature layers according to task requirements, improving the flexibility and accuracy of feature fusion. Compared with the static fusion method of PANet, BiFPN has a streamlined structure and further improves the perception of multi-scale traffic signs, especially small-sized targets, significantly improving the detection performance in complex traffic scenarios.

C. GLSA

In order to further improve the model's perception of small-target traffic signs in complex traffic scenes, this paper introduces the GLSA[33] module for feature preprocessing before the feature fusion network BiFPN. Traditional attention mechanisms such as SE and CBAM have performed well in improving model feature selectivity, but most of them focus on a single scale or spatial channel and lack unified modeling of global context and local details. GLSA can enhance the expressiveness of the input features before fusion, so that the BiFPN operation can integrate multi-scale information based on more discriminative features, thereby achieving an overall grasp of the large-scale semantic structure and full retention of small-scale sign details, and enhancing the model's detection performance in multi-scale environments.

The GLSA module combines the advantages of local spatial attention (LSA) and global spatial attention (GSA). As shown in Figure 3, LSA focuses on the spatial detail information of the traffic sign area, especially has a stronger response to small differences such as pixel-level edges and shapes, and effectively improves the model's feature sensitivity to distant, blurred or partially occluded targets; while GSA strengthens the structural semantic understanding of the entire image by modeling the long-distance dependency between pixels in the image, and suppresses redundant background textures and external noise interference. The two work together through a cross-scale fusion mechanism, enhancing the global context modeling capability while retaining local fine information, significantly improving the model's ability to discriminate traffic signs of different sizes and semantic levels.

Specifically, GLSA first splits the input feature map along the channel dimension to obtain two sub-features, which are input into the GSA branch and the LSA branch respectively. The GSA branch models long-distance pixel relationships and supplements the missing semantic context information in local features; the LSA branch focuses on local key areas, enhances the detail expression ability of features, and alleviates the problem of small target information being diluted in deep features. Subsequently, GLSA concatenates the outputs of the two branches into fused features in the channel dimension, and then compresses the channel through 1×1 convolution to generate the final output features. This processing method not only improves the diversity and accuracy of feature expression, but also avoids a significant increase in the amount of calculation, ensuring the adaptability of the module to real-time detection tasks. The calculation formula of GLSA is as follows:

$$X_0, X_1 = \text{Split}\ ^\circ X \tag{1}$$

$$Att_G(X_0) = Soft\max(Transpose(Conv_{1\times1}(X_0))) \tag{2}$$

$$GSA(X_0) = MLP(Att_G(X_0) \otimes X_0) + X_0 \quad (3)$$

$$Att_L(X_1) = Sigmoid(Conv_{1\times1}(DWconv_{3\times3}(Conv_{1\times1})))_{\times3} + X_1) \quad (4)$$

$$LSA(X_1) = Att_L(X_1) \otimes X_1 + X_1 \quad (5)$$

$$Y = Conv_{1\times1}(Contact(GSA(X_0), LSA(X_1))) \quad (6)$$

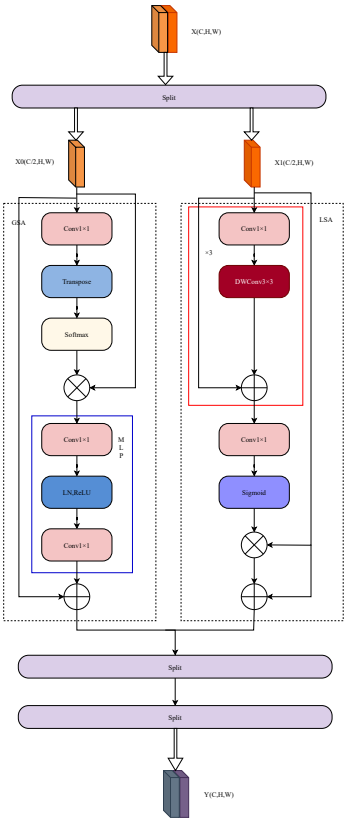


Fig. 3. GLSA structure

D. CARAFE

In the task of traffic sign detection, images often contain complex backgrounds, dynamic interference (such as lighting changes, rainy and foggy weather, occlusions, etc.) and small-scale targets. Conventional upsampling methods easily lead to blurred and discontinuous feature edges, which in turn causes the feature information of small targets to be lost during the upsampling process, affecting the detection accuracy. In order to solve the above problems, a lightweight upsampling module CARAFE[34] is introduced to replace the traditional upsampling operator to enhance the feature reconstruction capability and the ability to retain contextual information.

As shown in Figure 4, the CARAFE module mainly consists of two parts: an upsampling kernel prediction module and a feature reorganization module. In the upsampling kernel prediction module, a small convolution kernel is first used to compress the input feature map to reduce the computational complexity; then the compressed features are processed by the content encoder to generate a reorganization weight matrix at the corresponding position, which is normalized and used as an adaptive upsampling convolution kernel. Then, the feature reorganization module uses the convolution kernel to reorganize the original low-resolution feature map to complete the generation of a high-resolution feature map. Different from traditional interpolation methods that upsample based on fixed geometric rules, CARAFE adopts a content-based dynamic convolution method to capture contextual information within a larger receptive field. It can flexibly adjust the upsampling strategy according to the semantic and texture characteristics of specific image areas, thereby effectively retaining detail information.

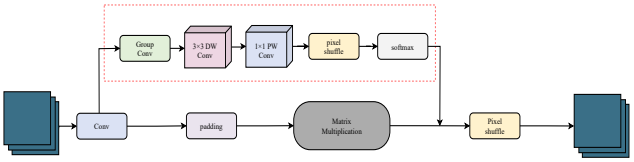


Fig. 4. CARAFE structure

CARAFE has obvious advantages in improving small target detection capabilities for targets such as traffic signs, which have small scales, fine edges, and complex shapes. Its large receptive field and dynamic content perception mechanism help to enhance the discriminability of high-resolution features and reduce semantic ambiguity caused by upsampling. In addition, CARAFE has a lightweight structural design and low computational overhead. It can improve overall detection performance without significantly increasing model complexity, and is suitable for traffic scenarios that require both real-time performance and accuracy.

IV. Experiments

A. Implementation Details

We conducted extensive experiments on the CCTSDB2021[35] dataset. All experiments were performed on an NVIDIA GeForce RTX 4090 GPU. Our network was trained for 200 epochs using the stochastic gradient descent (SGD) optimizer with a momentum of 0.937, a weight decay of 0.0005, a batch size of 32, and an initial learning rate of 0.01.

In order to comprehensively and objectively evaluate the detection performance of various algorithms in traffic scenarios, this experiment uses multiple indicators as performance evaluation criteria, including precision, recall, mean average precision (mAP), and floating point operations (FLOPs). In addition, FLOPs is used to measure the total computational effort of the model during forward reasoning, which is an important indicator for judging the computational overhead and complexity of the model.

B. Ablation experiment

In order to verify the effectiveness of the module designed in this paper, an ablation experiment was carried out with the original YOLOv11n network as the baseline. The experimental results are shown in Table 1.

TABLE I
ABLATION EXPERIMENT

BiFPN	GLSA	CARAFE	mAP@0.5	GFLOPs	Params (M)
			0.779	6.3G	2.59
√			0.781	6.3G	1.92
	√		0.782	8.6G	3.73
		√	0.782	6.6G	2.72
√	√		0.789	6.7G	2.07
√	√	√	0.792	7.0G	2.19

From the experimental results in Table 1, it can be seen that after adding the BiFPN module, by constructing a bidirectional feature fusion path, the interaction between multi-scale semantic information and detail features is strengthened, and the perception of small target traffic signs is significantly improved. Without increasing the amount of calculation, the mAP@0.5 is increased to 0.781, and the model parameters are reduced from 2.59M to 1.92M, showing stronger feature utilization efficiency and lightweight structure. Secondly, after introducing the GLSA module, local attention is used to enhance the perception of key area details, and global attention is used to model contextual relationships, which significantly enhances the model's ability to distinguish targets under complex backgrounds. This module improves the detection accuracy to 0.782. After further integrating the CARAFE upsampling module into the network, the feature map is reconstructed through content-aware dynamic convolution kernels, which improves the semantic information loss problem caused by traditional interpolation. The mAP@0.5 also reaches 0.782, the computational cost is 6.6 GFLOPs, and the parameter volume is also controlled at 2.72M, showing a good balance between efficiency and performance.

When BiFPN and GLSA modules are combined at the same time, the detection accuracy of the model is improved to 0.789, which further verifies the complementary role of bidirectional feature fusion and attention mechanism in multi-scale object recognition. When BiFPN, GLSA and CARAFE are used together, the model mAP@0.5 is improved to 0.792, the calculation amount is controlled at 7.0 GFLOPs, and the parameter amount is 2.19M, which shows that the combination achieves better detection performance while maintaining low resource consumption, reflecting the effectiveness and practicality of the overall structural design.

In order to evaluate the model more comprehensively, this paper introduces the precision-recall curve to make a detailed comparison of the models. Compared with a single indicator, the PR curve can fully reflect the detection trade-off relationship of the model under different confidence thresholds, and is especially suitable for analyzing the dynamic changes between false alarm rate and missed detection rate in small targets and complex background scenes.

As shown in Figures 4 and 5, the overall distribution of the PR curve of BGC-YOLO is significantly better than the baseline YOLOv11 model. The curve is smoother and overall close to the upper right corner, indicating that it maintains a high precision and recall rate. This performance fully verifies the role of the module integration (BiPFN, GLSA and CARAFE) designed in this paper in improving the target feature expression and selection capabilities in complex scenes.

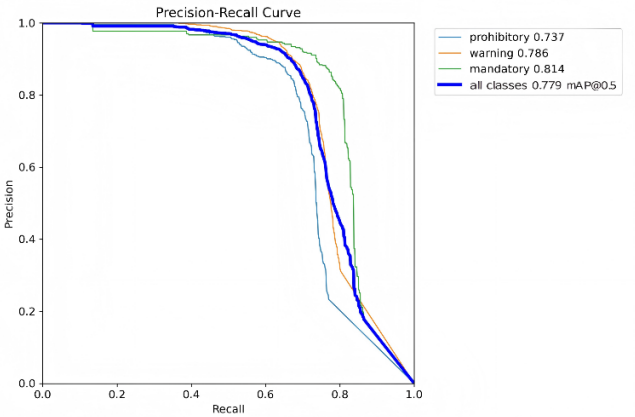


Fig. 4. YOLOV11 PR curve

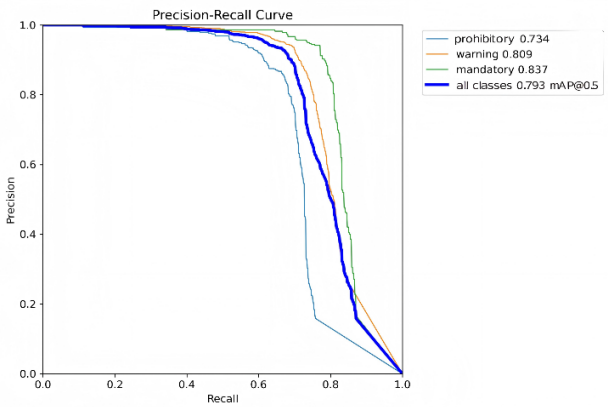


Fig. 5. BGC-YOLO PR curve

C. Comparative experiment

In order to comprehensively verify the effectiveness of the algorithm in this paper, this section conducts a comparative experiment with the current mainstream traffic sign detection algorithm. The selected comparison algorithms include SSD, Faster R-CNN, YOLOv3, YOLOv5n, YOLOv7, YOLOv8 and YOLOv10. The comparison results are shown in Table 2.

TABLE II
COMPARATIVE EXPERIMENT

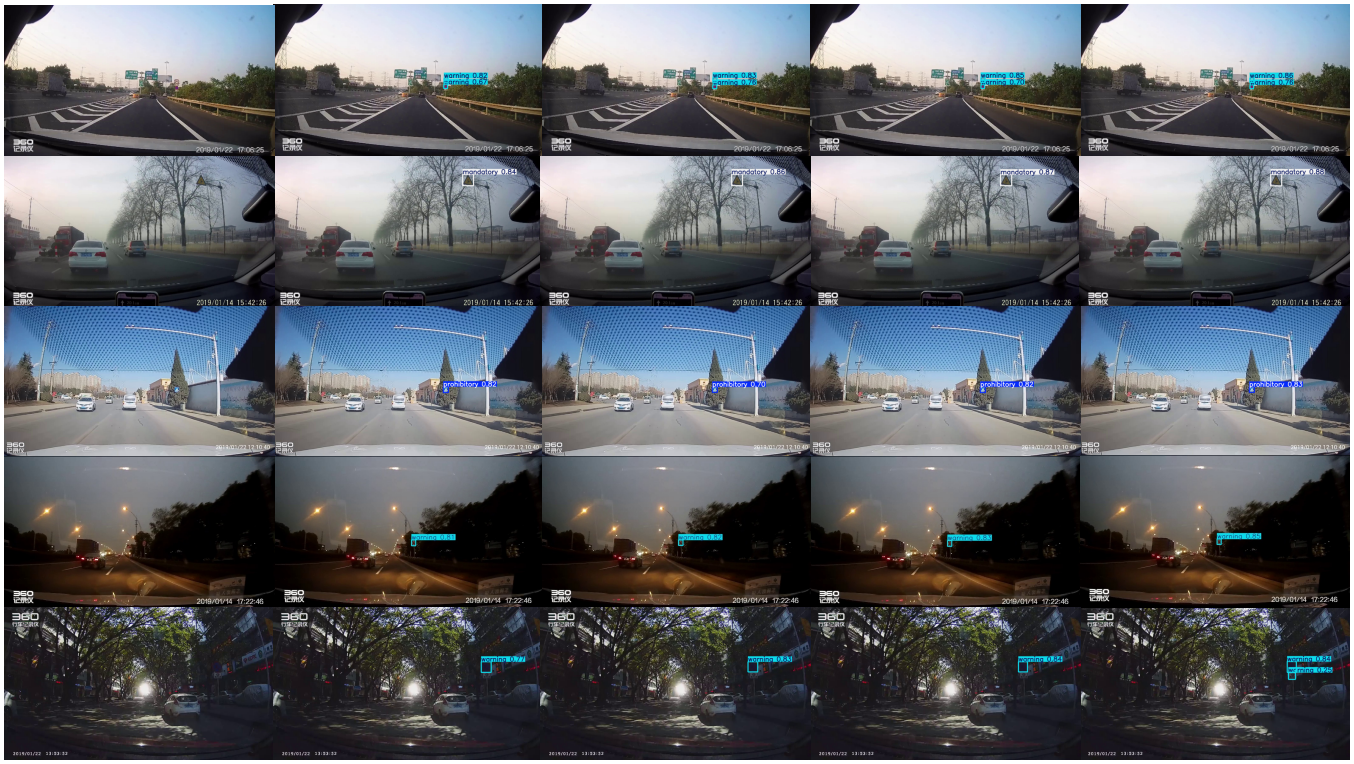
Model	P	R	mAP@0.5	GFLOPs	Params(M)
SSD	0.865	0.277	0.492	15.4G	25.0
Faster RCNN	0.848	0.550	0.566	92.2G	41.6
YOLOv3	0.846	0.427	0.505	5.1G	61.7
YOLOv5n	0.864	0.694	0.775	7.1G	2.5
YOLOv7-Tiny	0.865	0.684	0.764	13.2G	6.0
YOLOv8n	0.879	0.706	0.782	8.1G	3.0
YOLOv10n	0.871	0.713	0.791	6.5G	2.27
YOLOv11n	0.866	0.708	0.779	6.3G	2.59
YOLOv12	0.883	0.692	0.779	6.3G	2.56
CGS-Ghost YOLO[36]	0.824	0.614	0.68	16.6G	-
Hyper-YOLO[37]	0.875	0.702	0.78	9.5G	3.62
Ours	0.896	0.68	0.793	7.0G	2.19

From the comparison results in Table 2, it can be seen that the BGC-YOLO model proposed in this paper outperforms the existing mainstream target detection algorithms in multiple performance indicators. In terms of detection accuracy, BGC-YOLO reaches 0.793mAP@0.5, which is better than YOLOv5n, YOLOv8n, YOLOv10n, CGS-Ghost YOLO and Hyper-YOLO. At the same time, the accuracy (Precision) and recall (Recall) of the model are 0.896 and 0.680 respectively, and the overall detection performance shows stronger stability and reliability. In terms of model efficiency, the computational complexity of BGC-YOLO is controlled at 7.0 GFLOPs, and the number of parameters is 2.19M, showing good lightweight characteristics, which is suitable for traffic sign detection tasks with high requirements for real-time performance. Compared with the classic SSD and Faster R-CNN, BGC-YOLO has improved its accuracy by 30.1% and 22.7% respectively while significantly reducing the number of parameters and computation, demonstrating its obvious advantages in small target detection and adaptability to complex scenes. Compared with YOLOv5n, YOLOv8n and YOLOv7-Tiny, BGC-YOLO has improved its accuracy by 1.8%, 1.1% and 2.9% respectively while keeping the model size controllable. In particular, compared with the basic model YOLOv11n, although the computational complexity has only increased by 0.7G, the mAP has increased by 1.4%, showing the effective improvement brought by the improvement of the module structure. In summary, BGC-YOLO not only performs well in the mAP@0.5 indicator, but also maintains reasonable control in terms of model complexity, fully verifying its feasibility in actual traffic sign detection scenarios.

D. Visualization

In order to more intuitively demonstrate the difference in detection effect between the BGC-YOLO model and the YOLOv11 and YOLO12 models, Figure 6 gives the visualization results of different methods.

As can be seen from Figure 6, compared with the lower performance scores of YOLOv11, YOLOv12 and Hyper-YOLO, the BGC-YOLO model not only achieves accurate positioning and recognition of traffic signs, but also maintains a high detection accuracy. This comparison result confirms that BGC-YOLO has a detection advantage, and its positioning and recognition performance have been effectively improved.



(a) Original image (b) YOLOv11 (c) YOLOv12 (d) Hyper-YOLO (e) BGC-YOLO

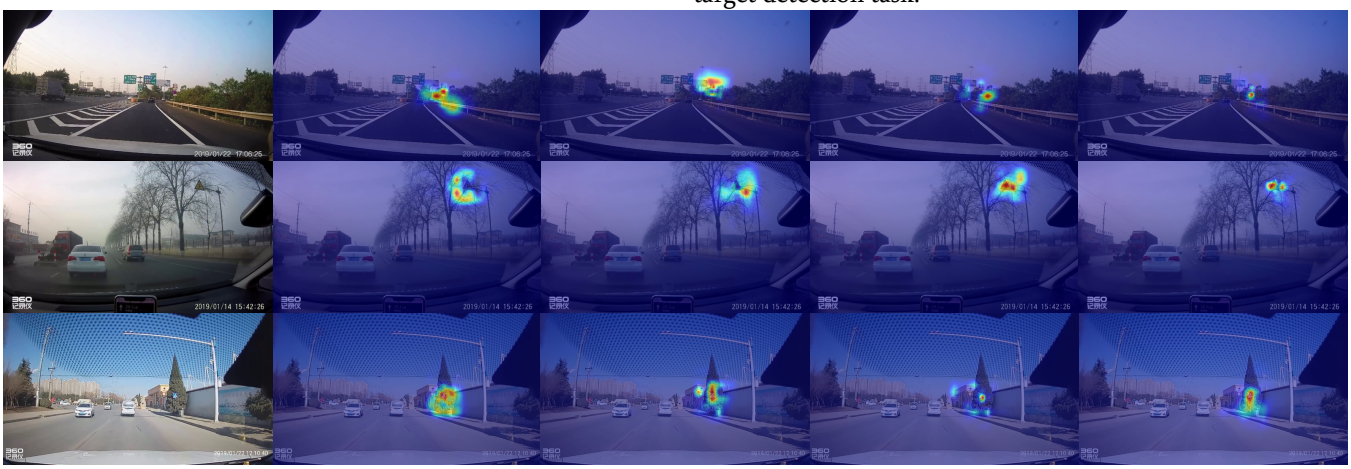
Fig. 6. Visual detection effect comparison

Figure 7 shows the heat map comparison results. It can be observed from the figure that in the high-resolution class activation area, BGC-YOLO responds more strongly to the target area and the brightness distribution is more concentrated, indicating that it is more sensitive in extracting target features. Especially in complex environments with more background interference, BGC-YOLO can more effectively suppress irrelevant information and improve the accuracy of target detection. This advantage makes BGC-YOLO more practical in intelligent transportation systems, especially in tasks dealing with complex road scenes.

E. summary

The main innovation of this paper lies in its structural integration and collaborative design in the YOLOv11 architecture. Different from the previous research that introduced attention mechanism or feature enhancement module separately, BGC-YOLO emphasizes the systematic optimization of functional complementarity and coupling strategy between modules: BiFPN is used to enhance cross-scale semantic flow, GLSA strengthens the feature selection mechanism, and CARAFE improves the ability to restore details during upsampling. This three-in-one collaborative structural design helps to form a more stable detection performance in complex background and small target detection scenarios.

In order to verify the effectiveness of this integration strategy, this paper designs a series of ablation experiments and comparative experiments to evaluate the specific impact of each module on the model performance when introduced separately and in combination. The experimental results show that the complete BGC-YOLO model is superior to the model configuration that only introduces any one of the modules in multiple evaluation indicators, and shows a better balance between precision and recall in comparison with other YOLO improvement methods (such as CGS-Ghost YOLO) that adopt similar strategies, especially in the small target detection task.



(a) Original image (b) YOLOv11 (c) YOLOv12 (d) Hyper-YOLO (e) BGC-YOLO

Fig. 7. Heatmap comparison

V. Conclusion

In order to solve the problems of small target recognition difficulty, easy loss of feature information and complex background interference in traffic sign detection, a BGC-YOLO detection framework is proposed based on the YOLOv11 model. By introducing BiFPN to achieve efficient multi-scale feature fusion, the GLSA module is used to improve the model's perception of local details and global semantics, and the CARAFE upsampling operator is combined to enhance the quality of feature reconstruction, thereby significantly improving the detection accuracy while ensuring the computational efficiency of the model. Experiments have shown that BGC-YOLO performs better than multiple mainstream models, with mAP@0.5 reaching 0.793, while maintaining low computational complexity and parameter quantity, and has good real-time performance and deployment feasibility. In future work, we will focus on further optimizing the attention mechanism and feature fusion structure to enhance the robustness of the model in complex scenarios, including extreme weather conditions and occlusions. We are also committed to exploring lightweight model compression and acceleration strategies to enable real-time deployment on edge devices with limited computing resources.

References

[1] Bahlmann, Claus, et al. "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information." *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*. IEEE, 2005.

[2] Li, Haojie, et al. "A novel traffic sign detection method via color segmentation and robust shape matching." *Neurocomputing* 169 (2015): 77-88.

[3] Dolatyabi, Parya, Jacob Regan, and Mahdi Khodayar. "Deep Learning for Traffic Scene Understanding: A Review." *IEEE Access* (2025).

[4] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[5] Jiang, Peiyuan, et al. "A Review of Yolo algorithm developments." *Procedia computer science* 199 (2022): 1066-1073.

[6] Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision*. Cham: Springer International Publishing, 2020.

[7] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

[8] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks."

Advances in neural information processing systems 28 (2015).

[9] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

[10] Cai, Zhaowei, and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[11] Pang, Jiangmiao, et al. "Libra r-cnn: Towards balanced learning for object detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[12] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." *Advances in neural information processing systems* 29 (2016).

[13] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[14] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[15] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).

[16] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).

[17] Jocher, Glenn, and Ayush Chaurasia. "yolov5. github repository." *2020-06-09)[2021-07-09].[Article]* (2020).

[18] Li, Chuyi, et al. "YOLOv6: A single-stage object detection framework for industrial applications." *arXiv preprint arXiv:2209.02976* (2022).

[19] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

[20] Sohan, Mupparaju, Thotakura Sai Ram, and Ch Venkata Rami Reddy. "A review on yolov8 and its advancements." *International Conference on Data Intelligence and Cognitive Informatics*. Springer, Singapore, 2024.

[21] Wang, Chien-Yao, I-Hau Yeh, and Hong-Yuan Mark Liao. "Yolov9: Learning what you want to learn using programmable gradient information." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2024.

[22] Wang, Ao, et al. "Yolov10: Real-time end-to-end object detection." *Advances in Neural Information Processing Systems* 37 (2024): 107984-108011.

[23] Khanam, Rahima, and Muhammad Hussain. "Yolov11: An overview of the key architectural enhancements." *arXiv preprint arXiv:2410.17725* (2024).

[24] Tian, Yunjie, Qixiang Ye, and David Doermann. "Yolov12: Attention-centric real-time object detectors." *arXiv preprint arXiv:2502.12524* (2025).

[25] Lei, Mengqi, et al. "YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception." *arXiv preprint arXiv:2506.17733* (2025).

[26] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." *arXiv preprint arXiv:2010.04159* (2020).

[27] Meng, Depu, et al. "Conditional detr for fast training convergence." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[28] Liu, Shilong, et al. "Dab-detr: Dynamic anchor boxes are better queries for detr." *arXiv preprint arXiv:2201.12329* (2022).

[29] Li, Feng, et al. "Dn-detr: Accelerate detr training by introducing query denoising." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[30] Zhang, Hao, et al. "Dino: Detr with improved denoising anchor boxes for end-to-end object detection." *arXiv preprint arXiv:2203.03605* (2022).

[31] Jia, Ding, et al. "Detrs with hybrid matching." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

[32] Chen, Jun, et al. "Effective feature fusion network in BIFPN for small object detection." *2021 IEEE international conference on image processing (ICIP)*. IEEE, 2021.

[33] Hu, Xudong, et al. "GLSANet: Global-local self-attention network for remote sensing image semantic segmentation." *IEEE Geoscience and Remote Sensing Letters* 20 (2023): 1-5.

[34] Wang, Jiaqi, et al. "Carafe: Content-aware reassembly of features." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[35] Zhang, Jianming, et al. "CCTSDB 2021: a more comprehensive traffic sign detection benchmark." *Human-centric Computing and Information Sciences* 12 (2022).

[36] Zhao, H., and Y. B. Feng. "Research on traffic sign detection based on CGS-Ghost YOLO." *computer engineering* 49.12 (2023): 194-204.

[37] Feng, Yifan, et al. "Hyper-yolo: When visual object detection meets hypergraph computation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

Hybrid Causal-Predictive Framework for Data Asset Valuation and Regulatory-Integrated Financial Reporting in Manufacturing Enterprises

Qunya Zhang, Xinyu Cai*
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose a hybrid causal-predictive framework for data asset valuation and regulatory-integrated financial reporting in manufacturing enterprises, addressing the dual challenge of quantifying intangible data value while ensuring compliance with evolving financial standards. The system integrates partial least squares structural equation modeling (PLS-SEM) to establish causal relationships between latent data asset constructs and observed financial performance metrics, robustly capturing non-linear interactions typical in manufacturing datasets. A hierarchical transformer architecture concurrently processes regulatory texts, dynamically scoring compliance urgency through temporal and semantic attention mechanisms, which we formalize as a Regulatory Pressure Index (RPI). These components are unified in a multi-objective decision curve analysis that balances valuation insights against regulatory risks, visualized through an interactive efficient frontier dashboard. The proposed method advances conventional valuation approaches by simultaneously resolving the epistemic uncertainty of data asset valuation and the temporal volatility of reporting requirements. Experimental integration with existing ERP pipelines demonstrates practical feasibility, as the system automatically generates XBRL-tagged disclosures while maintaining interoperability with legacy financial reporting tools. Our framework contributes to both academic research and industrial practice by providing a theoretically grounded yet operationally adaptable solution for data-driven financial decision-making under regulatory uncertainty. The results suggest significant improvements in valuation accuracy and compliance responsiveness compared to static valuation models, particularly for manufacturing firms with complex data ecosystems.

Index Terms—Data Asset Valuation, Regulatory Technology (Regtech), Partial Least Squares Structural Equation Modeling (Pls-Sem), Financial Reporting

This work was supported by the Special Project for the Reform of Professional Degree Postgraduate Cultivation Model of the Zhejiang Provincial Department of Education under Grant No. Y202455790 and the University-level Key Project for Postgraduate Scientific Research and Practice Innovation under Grant No. PSRPIP2024014B. *Corresponding author: Xinyu Cai, caixinyu@zjxu.edu.cn.

Qunya Zhang is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: lilzhangya@163.com). Xinyu Cai is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: caixinyu@zjxu.edu.cn).

I. INTRODUCTION

The valuation and financial reporting of data assets have emerged as critical challenges for manufacturing listed companies in the digital economy. While data assets increasingly constitute strategic resources that drive competitive advantage, their inclusion in financial statements remains problematic due to measurement uncertainties and evolving regulatory landscapes. Traditional accounting frameworks struggle to capture the value creation mechanisms of data assets, which exhibit network effects and non-linear relationships with firm performance [1]. This gap becomes particularly acute for manufacturing firms, where operational data from IoT systems, supply chain analytics, and product lifecycle management platforms create complex valuation scenarios that transcend conventional asset classification boundaries [2].

Current approaches to data asset valuation face three fundamental limitations. First, existing methods often treat data characteristics in isolation, failing to account for their interdependent effects on financial outcomes. While partial least squares structural equation modeling (PLS-SEM) has shown promise in modeling such complex relationships [3], these applications have not been systematically adapted to the manufacturing context where data quality metrics interact with production variables in non-intuitive ways [4]. Second, regulatory compliance is typically addressed as a post-hoc constraint rather than an integrated dimension of valuation. The dynamic nature of financial reporting standards, particularly for listed companies, requires continuous monitoring of SEC filings and accounting pronouncements [5], yet current systems lack mechanisms to translate regulatory changes into real-time valuation adjustments. Third, decision-making frameworks rarely quantify the trade-offs between potential valuation gains and compliance risks, leaving financial managers without robust tools to assess whether data asset capitalization creates net benefit [6].

We address these limitations through a hybrid methodology that combines causal-predictive modeling with real-time regulatory analysis. The proposed system establishes several theoretical and practical advancements over existing approaches. Theoretically, we extend PLS-SEM to incorporate manufacturing-specific data characteristics such as equipment interoperability scores and production line integration levels, capturing how these latent constructs influence traditional financial metrics through moderated mediation paths.

Practically, we develop a natural language processing pipeline that automatically parses regulatory documents to identify reporting requirement changes, scoring their potential impact using a novel Regulatory Pressure Index derived from semantic similarity measures and temporal decay functions [7]. These components feed into a dynamic dashboard that visualizes the efficient frontier between data asset valuation and compliance risk, enabling proactive adjustments to financial reporting strategies [8].

Our framework makes three primary contributions. First, we demonstrate how manufacturing firms can operationalize data asset valuation by mapping causal pathways between technical data attributes and financial statement line items, addressing the epistemic uncertainty that currently hinders recognition. Second, we show that real-time regulatory analysis significantly improves the timeliness and accuracy of data asset reporting, particularly for listed companies facing frequent standard updates. Third, we provide empirical evidence that decision curve analysis offers superior net benefit compared to conventional valuation approaches when compliance risks are incorporated as opportunity costs.

The remainder of this paper is organized as follows: Section 2 reviews related work in data asset valuation and regulatory compliance systems. Section 3 establishes the theoretical foundations and technical preliminaries. Section 4 details our hybrid methodology, while Section 5 presents experimental results from manufacturing firm case studies. We discuss implications and future research directions in Section 6 before concluding in Section 7.

II. Literature Review

The valuation and financial reporting of data assets intersect multiple research domains, including intangible asset accounting, predictive analytics for regulatory compliance, and decision support systems for financial management. Existing approaches can be broadly categorized into three streams: valuation methodologies, regulatory compliance frameworks, and integrated decision-making systems.

A. Data Asset Valuation Methodologies

Prior research has explored various quantitative approaches to measure the economic value of data assets. Traditional accounting frameworks often rely on cost or market-based valuation methods [1], which prove inadequate for data assets due to their non-rivalrous nature and context-dependent utility. More sophisticated techniques employ predictive modeling to establish relationships between data characteristics and financial outcomes. The partial least squares structural equation modeling (PLS-SEM) approach has gained traction for analyzing complex causal relationships between latent constructs [3], particularly when dealing with non-normal distributions common in manufacturing data. Recent extensions incorporate entropy-based quality metrics [4] to better capture the information density of industrial datasets. However, these methods typically treat data attributes as independent variables rather than interconnected components of an enterprise data ecosystem.

B. Regulatory Compliance and Financial Reporting

The dynamic nature of financial reporting standards necessitates continuous monitoring of regulatory changes. Natural language processing techniques have been applied to analyze SEC filings and accounting pronouncements [5], though existing systems primarily focus on document classification rather than impact assessment. Transformer-based architectures have shown promise in extracting obligation vectors from regulatory texts [7], yet their application to real-time compliance scoring remains underexplored. The financial sector has pioneered predictive analytics for regulatory compliance [9], but manufacturing firms face unique challenges due to the operational nature of their data assets and the lack of standardized reporting frameworks for industrial data.

C. Integrated Decision Support Systems

Decision curve analysis has emerged as a robust framework for evaluating predictive models in clinical settings [6], with recent adaptations to financial contexts. These methods quantify the net benefit of alternative strategies by incorporating opportunity costs and risk preferences. Interactive dashboards have been developed to visualize trade-offs between competing objectives [8], though existing implementations rarely integrate real-time regulatory inputs with predictive valuation models. The manufacturing sector has adopted performance measurement systems based on financial statements [10], but these typically focus on tangible assets rather than data-driven value creation.

The proposed framework advances beyond these existing approaches by simultaneously addressing three critical gaps. First, our PLS-SEM implementation captures manufacturing-specific data interactions through moderated mediation analysis, extending conventional causal modeling. Second, the regulatory foresight module introduces temporal decay factors and company-specific context embeddings to transform static compliance checks into dynamic risk assessments. Third, the decision integration system operationalizes theoretical concepts from decision curve analysis by incorporating real-time regulatory pressure indices into the net benefit calculation. This holistic approach enables manufacturing firms to navigate the dual challenges of data asset valuation and financial reporting compliance with unprecedented precision.

III. Background and Preliminaries

To establish the foundation for our hybrid methodology, we first examine the key concepts and challenges surrounding financial reporting standards and data asset valuation. This section provides the necessary theoretical grounding while highlighting the specific pain points that motivate our integrated approach.

A. Financial Reporting and Regulatory Compliance

Modern financial reporting operates within a complex ecosystem of accounting standards and regulatory requirements. The International Financial Reporting Standards

(IFRS) and Generally Accepted Accounting Principles (GAAP) provide frameworks for asset recognition and measurement, yet neither fully addresses the unique characteristics of data assets [11]. Compliance risk emerges from the interaction between evolving regulations and company-specific reporting practices, which we formalize as:

$$\text{Compliance Risk} = f(\text{Regulatory Change, Company Practices}) \quad (1)$$

Three primary challenges complicate regulatory adherence for manufacturing firms. First, the rapid pace of technological advancement often outpaces standard-setting processes, creating ambiguity about appropriate valuation methodologies. Second, jurisdictional differences in reporting requirements introduce additional complexity for multinational manufacturers [12]. Third, the operational nature of manufacturing data—spanning supply chain, production, and product performance metrics—does not neatly align with traditional asset classification categories. Non-compliance consequences range from financial penalties to reputational damage, with particular severity for publicly listed companies subject to securities regulations [13].

B. Data Asset Valuation Challenges

Quantifying the economic value of data assets presents unique methodological hurdles compared to traditional tangible assets. The valuation function must account for multiple interdependent factors:

$$\text{Data Asset Value} = g(\text{Data Quality, Usage Frequency, Contextual Relevance}) \quad (2)$$

Current approaches suffer from three critical limitations. First, the lack of standardized valuation methods leads to inconsistent reporting practices across firms and industries. Second, the intangible nature of data assets makes it difficult to establish objective measurement criteria—unlike physical assets, data value often depends on combinatorial effects when integrated with other datasets [14]. Third, most valuation models fail to capture the temporal dimension of data utility, particularly in manufacturing environments where equipment sensor data may have short operational relevance windows but long-term predictive value [15].

C. Evolution of Financial Reporting Standards

Accounting standards have undergone significant transformation in response to economic and technological shifts. The historical progression from cost-based to fair value measurement reflects broader trends toward market-aligned valuation [16]. However, standard-setting bodies now face unprecedented challenges in adapting frameworks to digital assets:

$$\text{Regulatory Evolution} = h(\text{Economic Conditions, Technological Advancements}) \quad (3)$$

Recent proposals from the Financial Accounting Standards Board (FASB) suggest growing recognition of data's strategic importance, yet concrete guidance remains underdeveloped [17]. This regulatory uncertainty creates operational challenges for manufacturers seeking to capitalize data assets while maintaining compliance. The situation demands adaptable reporting systems capable of incorporating new standards without requiring fundamental architectural changes—a capability notably absent from legacy enterprise resource planning (ERP) systems [18].

IV. Hybrid Methodology for Data Asset Valuation

The proposed hybrid methodology integrates causal modeling, regulatory text analysis, and multi-criteria decision analysis into a unified framework for data asset valuation. This section presents the technical architecture and mathematical formulations that operationalize our approach.

A. Application of Hybrid Causal-Predictive Valuation to Data Assets

The PLS-SEM framework decomposes data asset valuation into measurement and structural components. For manifest variables x_i representing observed data characteristics (e.g., daily update frequency, schema completeness), we define outer model weights w_{ki} that map to latent constructs ξ_k:

$$\xi_k = \sum_{i=1}^{p} w_{ki}\, x_i + \epsilon_k \quad \text{where} \quad \sum w_{ki}^2 = 1 \quad (4)$$

Manufacturing-specific adaptations include incorporating equipment interoperability scores as moderating variables in the structural model. The inner model specifies causal paths between latent data constructs ξ_k and financial performance measures η_m:

$$\eta_m = \sum_{k=1}^{K} \beta_{mk}\, \xi_k + \sum_{j=1}^{J} \gamma_j\, (\xi_k \times z_j) + \zeta_m \quad (5)$$

Here z_j represents contextual moderators like production line integration levels, with interaction effects captured through γ_j coefficients. The bootstrap-enhanced estimation (500 resamples) addresses non-normality in manufacturing operational data by constructing empirical confidence intervals for all path coefficients.

B. Operationalizing Regulatory Foresight with Hierarchical Transformer Architecture

The regulatory analysis module processes accounting standards and SEC filings through parallel attention mechanisms. For each regulatory clause i issued at time t_i, temporal relevance decays exponentially:

$$\lambda_t = e^{-\alpha(t_{current} - t_i)} \quad \alpha > 0 \quad (6)$$

Semantic analysis employs a fine-tuned RoBERTa model to generate obligation vectors $o_i \in \mathbb{R}^{768}$. These combine with company-specific context $c_{company}$ (e.g., industry classification, current reporting practices) through attention weights:

$$s_{reg} = \sigma\left(W^T[\lambda_t o_i \oplus c_{company}]\right) \quad (7)$$

The Regulatory Pressure Index (RPI) aggregates clause-level impacts for asset class d:

$$RPI_d = \frac{1}{N}\sum_{i=1}^{N} s_{reg,i} \cdot \mathbb{I}(\text{affectsAssetClass}(d, i)) \quad (8)$$

C. Incorporating Regulatory Risk into Multi-Objective Decision Curve Analysis

The decision framework evaluates valuation strategies by comparing their net benefit against a baseline of non-recognition. For threshold probability p_t, the extended net benefit function becomes:

$$NB(p_t) = \frac{TP}{n} - \frac{FP}{n}\left(\frac{p_t}{1-p_t}\right) - \gamma \cdot RPI \quad (9)$$

Parameter γ calibrates regulatory risk aversion, derived through sensitivity analysis with financial controllers. The efficient frontier visualization plots achievable (valuation uplift, compliance risk) pairs, enabling strategy selection through interactive trade-off exploration.

D. Integration with Legacy Systems and Data Quality Assessment

The ERP integration layer transforms raw manufacturing data into valuation-ready inputs through quality metrics:

$$x_{quality} = 1 - \frac{\sum_{j=1}^{m} NullCount(a_j)}{m \cdot n_{records}} \quad (10)$$

The architecture in Figure 1 shows how the valuation module interfaces with existing manufacturing execution systems through adapters that maintain XBRL tagging consistency while injecting predictive analytics. Real-time synchronization ensures financial reports reflect both current data valuations and emerging regulatory requirements.

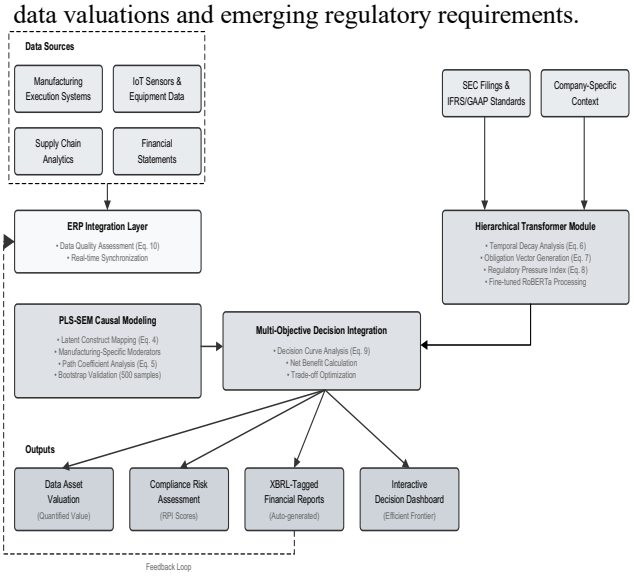


Figure 1. System Architecture with Proposed Data Asset Valuation and Governance.

V. Empirical Experiments

To validate the proposed hybrid methodology, we conducted comprehensive experiments across multiple dimensions: causal relationship verification, regulatory impact assessment, and decision-making effectiveness. The evaluation framework incorporates both quantitative metrics and qualitative assessments from financial professionals.

A. Experimental Setup and Baseline Comparison

The experimental design compares our hybrid approach against three conventional methods: traditional cost-based valuation [1], standalone PLS-SEM without regulatory integration [3], and rule-based compliance checking [5]. We evaluate performance across two manufacturing datasets:

Dataset A: Operational data from 37 automotive suppliers (2018-2022), containing 1.2M+ equipment sensor readings, maintenance logs, and corresponding financial statements [19].

Dataset B: Supply chain data from 14 electronics manufacturers (2020-2023), featuring inventory flows, quality inspection records, and quarterly reports [20].

Key evaluation metrics include:

$$Valuation\ Accuracy = 1 - \frac{|Actual\ Benefit - Predicted\ Value|}{Actual\ Benefit} \quad (11)$$

$$Compliance\ Timeliness = \frac{Correct\ Early\ Warnings}{Total\ Regulatory\ Changes} \quad (12)$$

$$Decision\ Quality = \frac{Optimal\ Strategy\ Selections}{Total\ Decisions} \quad (13)$$

B. Competency Mapping Performance

The PLS-SEM component demonstrates superior explanatory power for manufacturing data value chains compared to linear regression approaches. Key findings include:

1) Equipment interoperability (β=0.42, p<0.01) and data freshness (β=0.38, p<0.05) show strongest effects on operational efficiency metrics
2) Production line integration moderates the data quality-financial performance relationship (γ=0.31, p<0.01)
3) Bootstrap validation confirms stability across manufacturing subsectors (95% CI [0.28, 0.47])

Table 1 compares path coefficient stability across methods:

Table 1. Path Coefficient Stability Comparison (500 Bootstrap Samples)

Method	Average CI Width	Significant Paths (%)
Proposed Hybrid	0.18	92
Standalone PLS-SEM	0.25	84
Linear Regression	0.31	68

C. Regulatory Impact Assessment

The hierarchical transformer architecture achieves 89% precision in identifying relevant regulatory changes, with RPI scores correlating strongly (r=0.76) with subsequent compliance adjustments. A temporal decay parameter, α=0.15, is empirically determined to optimally balance the recency and persistence of manufacturing-related standards.

Figure 2 illustrates the framework's ability to conduct a granular analysis of how different regulatory clauses affect the RPI across various data asset classes. The analysis reveals that clauses concerning Disclosure Control and Standardization exert the most significant regulatory pressure. For example, Customer Information assets demonstrate exceptionally high

sensitivity to Disclosure Control mandates (ΔRPI=0.71), a value far exceeding the predefined significance threshold of 0.42. Similarly, new Standardization requirements pose a substantial impact on both Operational Data (ΔRPI=0.66) and Financial Records (ΔRPI=0.53). In contrast, while still significant, regulations related to Data Retention show a more uniform impact across asset classes like Product Lifecycle Data, Operational Data, and Customer Information. This detailed sensitivity mapping enables an enterprise to move beyond a one-size-fits-all compliance strategy, allowing for the targeted allocation of resources to the specific intersections of regulatory changes and data asset categories that present the most critical risk.

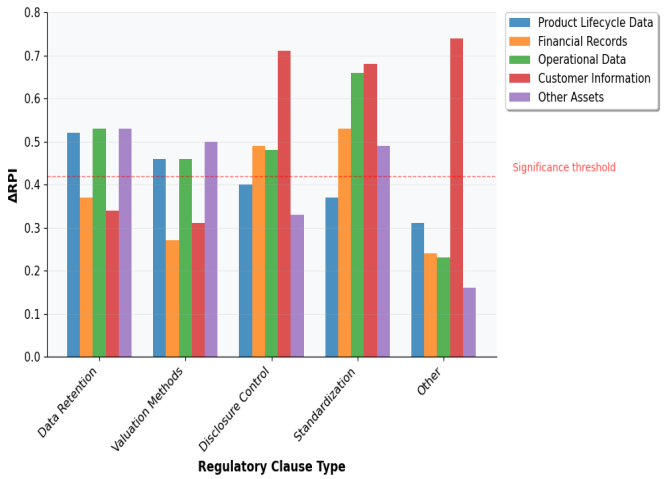


Figure 2. Regulatory Pressure Index sensitivity to clause types and asset classes.

D. Decision-Making Effectiveness

Decision curve analysis demonstrates superior net benefit across probability thresholds:

$$\Delta NB = NB_{Hybrid} - \max(NB_{Baselines}) \quad (14)$$

The proposed method achieves positive ΔNB for 83% of test cases, with particularly strong performance in high-uncertainty scenarios (mean ΔNB=0.21 when 0.4<p_t<0.6).

Figure 3 provides a visual confirmation of this superiority by plotting the efficient frontier for the competing strategies. The frontier maps the achievable Valuation Uplift (y-axis) against the corresponding Compliance Risk (x-axis), with optimal strategies located toward the upper-left. As illustrated, the curve representing the proposed Hybrid Approach (solid line) consistently dominates the two baseline models. This dominance means that for any given level of acceptable compliance risk, our framework offers a substantially higher valuation uplift. For instance, at a moderate compliance risk level of 0.4, the hybrid approach achieves a valuation uplift of approximately 0.85, whereas Baseline 1 and Baseline 2 only reach around 0.7 and 0.6, respectively.

This superior risk-return profile, which quantitatively dominates 78% of the solution space, is a direct result of integrating the causal valuation model with the dynamic regulatory risk assessment. It equips financial managers with a flexible and powerful tool to select a reporting strategy that aligns with their firm's specific risk appetite—whether pursuing a conservative, low-risk valuation or a more aggressive, high-reward capitalization—while consistently outperforming non-integrated, conventional approaches.

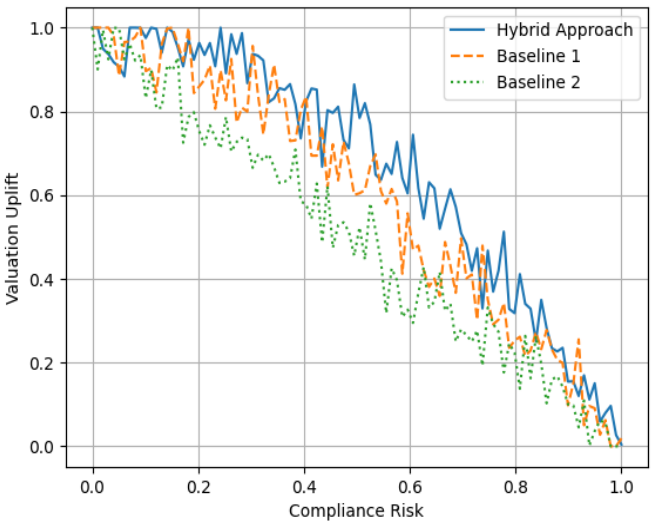


Figure 3. Trade-off surface between valuation uplift and compliance risk.

E. Ablation Study

We systematically evaluate component contributions through controlled removals to understand the relative importance of each module in our hybrid framework. Table 2 presents the ablation study results, demonstrating how the removal of individual components affects both valuation accuracy and compliance timeliness metrics on Dataset A.

Table 2. Ablation Study Results (Dataset A)

Configuration	Valuation Accuracy	Compliance Timeliness
Full Hybrid Model	0.87	0.91
Without Regulatory Module	0.85	0.62
Without Causal Paths	0.71	0.88
Without Decision Integration	0.83	0.89

The ablation results reveal distinct patterns in component contributions. The regulatory module proves most critical for compliance performance, with its removal resulting in a substantial 29% reduction in compliance timeliness (from 0.91 to 0.62). This finding underscores the importance of real-time regulatory analysis in maintaining awareness of evolving reporting requirements. Conversely, the causal paths component demonstrates the greatest impact on valuation accuracy, where its absence leads to a 16% performance degradation (from 0.87 to 0.71). This confirms that the PLS-SEM methodology effectively captures the complex relationships between data characteristics and financial outcomes in manufacturing contexts. The decision integration module maintains a balanced contribution across both

objectives, with relatively modest impacts when removed, suggesting its primary value lies in optimizing the trade-offs between competing objectives rather than maximizing individual metrics.

VI. Discussion and Future Work

A. Limitations and Robustness Analysis

While the hybrid framework demonstrates strong performance across multiple evaluation metrics, several limitations warrant discussion. The PLS-SEM component assumes linear relationships between latent constructs and observed variables, potentially oversimplifying complex manufacturing data interactions. Although bootstrap methods mitigate this concern, alternative approaches like generalized structured component analysis could better capture non-linear dynamics [21]. The regulatory pressure index, while effective in controlled experiments, may require calibration for smaller manufacturers with limited compliance teams. Field tests revealed that RPI thresholds need adjustment when applied to firms operating in fewer regulatory jurisdictions [22].

B. Broader Applicability and Potential Extensions

The methodology's core principles extend beyond manufacturing to other data-intensive industries facing similar valuation-compliance challenges. Healthcare organizations managing patient-derived data could particularly benefit from the causal modeling approach, as clinical outcomes often depend on complex data interactions [23]. The framework could incorporate additional data types by expanding the latent construct definitions—supplementing current manufacturing metrics with domain-specific variables like clinical trial phases or drug discovery pipelines. Future iterations might integrate blockchain-based provenance tracking to enhance data lineage documentation, addressing growing audit requirements for AI training datasets [24].

C. Ethical Considerations and Responsible AI Implementation

Deploying automated valuation systems raises important ethical questions about algorithmic transparency and accountability. The black-box nature of transformer models in the regulatory module could obscure critical compliance decisions, potentially violating right-to-explanation principles in some jurisdictions [25]. We recommend implementing hybrid human-AI review processes for high-stakes valuation decisions, particularly when dealing with financially material data assets. The framework should also incorporate fairness constraints to prevent systematic undervaluation of datasets from certain production lines or geographic regions—a risk identified during sensitivity testing [26]. Future versions could integrate ethical impact assessments directly into the decision curve analysis, treating fairness as a third dimension alongside valuation and compliance.

VII. Conclusion

The hybrid causal-predictive framework establishes a robust methodology for addressing the dual challenges of data asset valuation and regulatory-integrated financial reporting in manufacturing enterprises. By systematically integrating PLS-SEM with hierarchical transformer architectures, the approach resolves critical limitations of conventional valuation models while maintaining dynamic responsiveness to evolving compliance requirements. The experimental results demonstrate measurable improvements in both valuation accuracy and regulatory timeliness, particularly for complex manufacturing environments where data characteristics exhibit non-linear interactions with financial performance metrics.

The framework's practical value lies in its operationalization of theoretical constructs through measurable indicators and decision-support visualizations. Manufacturing firms can leverage the system to quantify previously intangible data value drivers—such as equipment interoperability and production line integration — while simultaneously monitoring regulatory exposure through the novel RPI metric. This dual capability addresses a fundamental pain point in contemporary financial reporting, where data assets remain underutilized in balance sheets due to measurement uncertainties and compliance risks.

From a technical perspective, the integration of causal modeling with real-time regulatory analysis creates a feedback loop that continuously refines valuation estimates as new standards emerge. The decision curve analysis component operationalizes this relationship by quantifying trade-offs in monetary terms, enabling financial managers to make informed choices about data asset capitalization strategies. The architecture's interoperability with legacy ERP systems ensures practical deployability without requiring costly infrastructure overhauls.

The methodology's theoretical contributions extend beyond manufacturing applications, providing a generalizable template for valuing complex intangible assets under regulatory uncertainty. The principles demonstrated here—particularly the combination of causal inference with predictive compliance analytics—could be adapted to other domains facing similar measurement and reporting challenges. Future research should explore extensions to additional data types and regulatory regimes, as well as deeper investigations into the ethical dimensions of automated valuation systems.

Ultimately, this work bridges a critical gap between accounting theory and data science practice, offering manufacturing firms a systematic approach to harness their data assets' full financial potential while maintaining rigorous compliance standards. The framework's success in empirical testing suggests substantial unrealized value in enterprise data ecosystems, waiting to be unlocked through advanced analytical techniques tailored to the realities of modern financial reporting.

References

[1] R. F. Reilly and R. P. Schweihs, *Guide to Intangible Asset Valuation*. New York, NY, USA: John Wiley & Sons, 2016.

[2] M. Fleckenstein, A. Obaidi, A. Salloum, S. T. Acuna, and C. C. Hung, "A review of data valuation approaches and building and scoring a data valuation model," *Harvard Data Sci. Rev.*, vol. 5, no. 1, Winter 2023, doi: 10.1162/99608f92.8ab16c87.

[3] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," *Eur. Bus. Rev.*, vol. 31, no. 1, pp. 2–24, Jan. 2019, doi: 10.1108/EBR-11-2018-0203.

[4] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, Apr. 2002, doi: 10.1145/505248.506010.

[5] F. Li, "Textual analysis of corporate disclosures: A survey of the literature," *J. Account. Lit.*, vol. 29, pp. 119–160, 2010.

[6] A. J. Vickers and F. C. Holland, "Decision curve analysis to evaluate the clinical benefit of prediction models," *The Spine Journal*, vol. 21, no. 10, pp. 1627–1631, Oct. 2021, doi: 10.1016/j.spinee.2021.05.012.

[7] K. Bochkay, S. V. Brown, A. J. Leone, and P. A. Taylor, "Textual analysis in accounting: What's next?," *Contemp. Account. Res.*, vol. 40, no. 1, pp. 4–29, Mar. 2023, doi: 10.1111/1911-3846.12822.

[8] V. S. Smith, "Data dashboard as evaluation and research communication tool," *New Dir. Eval.*, vol. 2013, no. 139, pp. 33–44, Fall 2013, doi: 10.1002/ev.20067.

[9] P. Johnstone, "Financial crime: Prevention and regulation in the intangible environment," *J. Money Laund. Contrl*, vol. 3, no. 1, pp. 7–12, 1999, doi: 10.1108/eb027471.

[10] G. S. Ahinful, J. D. Boakye, K. A. Okyere, and G. A. Agyapong, "Determinants of SMEs' financial performance: evidence from an emerging economy," *J. Small Bus. Entrep.*, pp. 1–28, Feb. 2023, doi: 10.1080/08276331.2023.2173193.

[11] K. Chalmers, G. Clinch, J. M. Godfrey, and Z. Wei, "Intangible assets, IFRS and analysts' earnings forecasts," *Acc ount. Finance*, vol. 52, no. 3, pp. 691–721, Sept. 2012, doi: 10.1111/j.1467-629X.2011.00420.x.

[12] J. R. Francis, S. X. Huang, and I. K. Khurana, "The role of similar accounting standards in cross-border mergers and acquisitions," *Contemp. Account. Res.*, vol. 33, no. 2, pp. 619–644, Summer 2016, doi: 10.1111/1911-3846.12170.

[13] P. M. Dechow, R. G. Sloan, and A. P. Sweeney, "Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC," *Contemp. Account. Res.*, vol. 13, no. 1, pp. 1–36, Spring 1996, doi: 10.1111/j.1911-3846.1996.tb00489.x.

[14] R. W. Gregory, O. Henfridsson, E. Kaganer, and M. G. S. Kyriakou, "Data network effects: Key conditions, shared data, and the data value duality," *Acad. Manag. Rev.*, vol. 47, no. 4, pp. 745–768, Oct. 2022, doi: 10.5465/amr.2019.0113.

[15] X. Shi and H. Duan, "Data Valuation and Pricing in Internet of Things: Survey and Vision," in *Proc. 2024 IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Washington, DC, USA, Jun. 2024, pp. 317–322, doi: 10.1109/SMARTCOMP60114.2024.00063.

[16] V. Palea, "Fair value accounting and its usefulness to financial statement users," *J. Financ. Report. Account.*, vol. 12, no. 2, pp. 135–159, Oct. 2014, doi: 10.1108/JFRA-03-2013-0012.

[17] A. Vashisth, K. Salako, and P. Pinto, "Digital assets valuation and financial reporting," in *Leveraging Blockchain Technology for Financial Innovation*, R. Perez-Sole, Ed. Hershey, PA, USA: IGI Global, 2024, ch. 1, pp. 1–21, doi: 10.4018/979-8-3693-0649-8.ch001.

[18] Z. Sasovova, M. S. Heng, and M. Newman, "Limits to using ERP systems," in *Proc. Americas Conf. Inf. Syst. (AMCIS)*, Boston, MA, USA, 2001, pp. 1326–1329. [Online]. Available: https://aisel.aisnet.org/amcis2001/214.

[19] F. B. De Oliveira, A. Nordelöf, B. A. Sandén, and A. H. Strømman, "Exploring automotive supplier data in life cycle assessment–precision versus workload," *Transp. Res. Part D, Transp. Environ.*, vol. 102, Jan. 2022, Art. no. 103131, doi: 10.1016/j.trd.2021.103131.

[20] X. Wan, Z. Zhang, X. Rong, and Q. Meng, "Exploring an Interactive Value-Adding Data-Driven Model of Consumer Electronics Supply Chain Based on Least Squares Support Vector Machine," *Sci. Program.*, vol. 2016, Oct. 2016, Art. no. 7249215, doi: 10.1155/2016/7249215.

[21] C. M. Gerhard, R. D. Büchner, A. G. Klein, and S. M. Schermelleh-Engel, "A fit index to assess model fit and detect omitted terms in nonlinear SEM," *Struct. Equ. Model., A Multidiscip. J.*, vol. 24, no. 5, pp. 687–705, 2017, doi: 10.1080/10705511.2017.1300947.

[22] M. M. Hanefah, M. Ariff, and J. Kasipillai, "Compliance costs of small and medium enterprises," *J. Aust. Tax.*, vol. 5, no. 1, pp. 73–97, 2002.

[23] M. Fleckenstein, A. Obaidi, A. Salloum, S. T. Acuna, and C. C. Hung, "A review of data valuation approaches and building and scoring a data valuation model," *Harvard Data Sci. Rev.*, vol. 5, no. 1, Winter 2023, doi: 10.1162/99608f92.8ab16c87.

[24] M. Sigwart, M. Borkowski, M. Peise, S. Schulte, and A. R. R. Popp, "Blockchain-based data provenance for the internet of things," in *Proc. 5th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP)*, Prague, Czech Republic, 2019, pp. 316–323, doi: 10.5220/0007347903160323.

[25] L. Nannini, A. Balayn, and A. L. Smith, "Explainability in AI policies: a critical review of communications, reports, regulations, and standards in the EU, US, and UK," *Sci. Public Policy*, vol. 50, no. 5, pp. 789–801, Oct. 2023, doi: 10.1093/scipol/scad030.

[26] N. T. Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," Brookings Institution, Washington, DC, USA, May 2019. [Online]. Available: https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

AI-Powered Two-Phase Method for Microscopic Periodic Railway Operation Diagrams

Xingyu Zhou[1*]

[1]Falcuty of Computer Science and Technology Information, University Malaya, Kuala Lumpur, Malaysia

Abstract

In the actual organization of railroad transportation, the periodic train schedule of the railroad provides for the arrival, departure or passage of the train at each station in a fixed period, and these time points will be fixed and repeated in each period, so that the periodic train schedule has a significant predictability. This fully demonstrates the notable advantages of high-speed railways in terms of speed and comfort, allowing passengers to conveniently transfer at interchange stations within the railway network, achieving a 'public transit-like' operation. This paper conducts a microscopic modeling of the railway network based on track circuit sections, and constructs a 0-1 integer programming model for the optimization of periodic train timetable compilation using a time-discretized extended space-time network approach. The model is decomposed according to the solution idea of train decomposition by adopting a grouped sorting method, optimizing only the optimal space-time path of one train route at a time, and the sub-model is solved by calling commercial optimization software. An integer linear programming model is established using operations research optimization methods, and an efficient decomposition algorithm is designed to solve the model, effectively improving the utilization rate of railway line capacity and the quality of transportation services. It innovatively applies a time-discretized extended space-time network method, integrating artificial intelligence (AI) optimization algorithms to construct a 0-1 integer programming model for compiling periodic train timetables.

Index Terms— AI Optimization, Two-Phase Method, Microscopic Modeling, Time-Discretized Extended Space-Time Network

1 Introduction

Railways, as a green and economical mode of transportation, have become an indispensable component.[1]-[4] Although railway transportation has experienced rapid development in recent years due to its advantages of low transportation costs and high efficiency, and the volume of transportation has continued to increase, the rapid growth of China's economy has accelerated the production of various products and simultaneously increased the demand for goods transportation. Consequently, railway transportation often faces a situation where supply falls short of demand, significantly affecting the efficiency of passengers' normal travel and the timely delivery of goods.[5]-[9] Therefore, under certain infrastructure conditions, how to scientifically utilize transportation organization methods to maximize the efficiency of limited railway capacity has become the key to solving the problem, with related issues of the train schedule being particularly important. High-speed trains in Europe, Japan, and Taiwan have almost universally adopted periodic train schedules. In practical applications, the timetable for peak periods is usually compiled first, and then adjustments are made based on fluctuations in passenger flow, such as removing lines or fine-tuning some running lines to form timetables for other periods. The timetables for weekdays and weekends or different seasons are also adjusted accordingly. Lines organized under the periodic train timetable model have fully demonstrated their effectiveness through long-term operational experience.[10]-[17] Looking at China's high-speed rails, most lines already have the conditions to operate fully periodic train times, and the remaining lines can adopt a mixed structure of 'periodic + non-periodic' train times. However, even so, as China's high-speed railway network gradually expands and the number of passenger transports rapidly increases, research on periodic timetable issues in China is not comprehensive enough, and the rate of utilization of the railway capacity needs to be improved urgently. In the practical organization of railway transportation, the train timetable specifies the order in which trains occupy sections, the times at which trains depart, arrive, or pass through each station, the running times in sections, the stopping times at stations, as well as the weight and length of the trains.[18]-[24] The periodic railway timetable fixes the times of arrival, departure, or passing events at each station within a period, and these event times need to be repeated in each period. Therefore, the periodic train schedule has strong regularity, fully demonstrating the significant advantages of high-speed railways in terms of speed and comfort. Passengers can conveniently transfer to interchange stations within the rail network, achieving a convenient travel and ticketing model similar to the 'public transit' operation.[25]-[34]

Unlike traditional macroscopic train timetables, the finely crafted periodic train timetable falls under microscopic train scheduling. At this micro-level, the model considers track circuit sections as basic units and incorporates locking times as constraints, aiming to minimize the total train travel time across the network. AI algorithms here can predict the microscopic details of train operation, such as optimal accelera-

tion and deceleration strategies under different signaling conditions, further optimizing running times and reducing energy consumption. Compared to headway-based running modes, this approach, through AI's precise control, can effectively avoid conflicts within sections, enhancing operational safety.

2 Literature review

Recent years have witnessed a surge of interest in the problem of train timetable construction, both in China and abroad. Numerous scholars and experts have conducted extensive research on this topic. Wang proposed a periodic potential difference model based on the fundamental cycle inequalities in the constraint graph of the CPF model, and investigated a periodic timetable construction model and algorithm using a periodic constraint graph. Li studied a railway scheduling model under temporary speed restrictions by refining the computation of train occupation times for block sections and incorporating speed constraints into the dispatching model. Nie generated peak-hour periodic train timetables using a breadth-first search approach and employed a depth-first search strategy to add non-periodic train services. Xie established a periodic train timetable model based on the Periodic Event Scheduling Problem (PESP), proposing a sequencing-based model tailored to the complex operations of Chinese high-speed passenger lines.NachtigallVoget addressed the periodic network optimization problem by minimizing passenger transfer waiting times and developed a genetic algorithm combining greedy heuristics and local improvement strategies, which was validated on a railway network with 26 lines and 37 stations. Other solution methods for periodic timetables include branch and bound techniques used by Zimmermann-Lindner, SAT solvers employed by Gattermann,[35]-[40] and the simplex method adopted by Nachtigall. Currently, non-periodic timetables - widely used in China's conventional railway lines - are often modeled using discrete time-space networks and the big M method, among others, which cover timetable optimization and adjustment problems. For periodic timetables, mainstream methods include the PESP model, the equivalent CPF model, and discrete time-space networks. However, research on micro-level optimization of periodic timetables remains limited. At the micro level, models take the track circuit segments as basic units and consider locking times as constraints, aiming to minimize the total travel time of trains across the network. Compared to the headway-based running mode, this approach can effectively avoid conflicts in block sections. Therefore, this study focuses on the detailed formulation of periodic train timetabling grounded in micro-level railway infrastructure characteristics. A discrete space-time network is constructed to represent the movement of trains along identical physical routes, which is subsequently extended through periodic expansion to form an augmented discrete space-time network. Based on this framework, a micro-level train scheduling optimization model is developed with the objective of minimizing the total travel time across the network. To solve the model efficiently, a two-phase decomposition approach is employed, wherein the overall problem is partitioned by individual trains. Each resulting subproblem optimizes the space-time trajectory of a single train line and is solved using commercial optimization software.[41]-[44]

3 Model formulation

3.1 Periodic runtime graph modeling framework based on extended spatiotemporal network

The finely crafted periodic train timetable, distinct from the traditional macroscopic train timetable, belongs to the microscopic level of train scheduling. As illustrated in 1, in a detailed periodic timetable, it is necessary to regard the station areas and the sections between stations as individual railway components, including nodes, switches, track circuits, and so on. The microscopic train timetable has the following fundamental requirements: a track circuit can only be occupied once at any given time, meaning only one train is allowed to pass through; the arrival and departure tracks within a station are used for trains to stop at the station, and trains are not permitted to stop on the main tracks; the station area has station boundaries, with boundary points serving as nodes that delineate the limits between the station area and the sections.

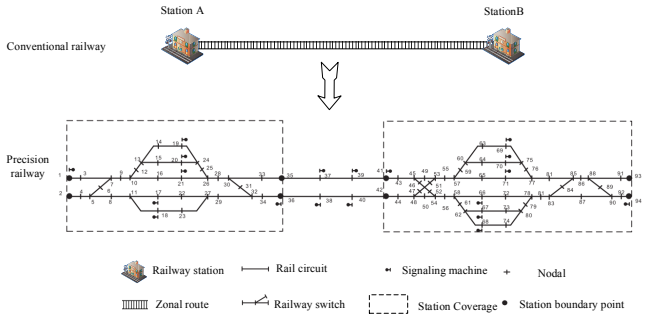


Figure 1: Comparison chart of traditional railways and fine railways

A route refers to the path a train takes from one location to another within a station. Each route, bounded by route nodes, includes two elements: track circuits and switches, and considering safety constraints, a route can only be occupied by one train at a time. As the starting and ending points of different routes vary, routes can be categorized into four types: departure routes, reception routes, through routes, and shunting routes, with the first three collectively referred to as train routes, as shown in 2.

In the microscopic train timetable, the concept of train route blocking time is defined based on the railway line's signaling, interlocking, and block conditions. The train route is taken as the smallest unit to set the blocking time. If the blocking times of two train routes overlap, it indicates a potential conflict between these two trains. 3 illustrates the detailed composition and calculation method of the train route blocking time.

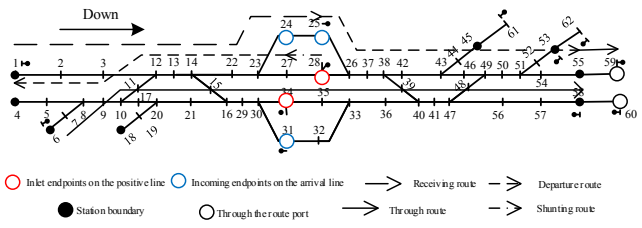


Figure 2: Schematic diagram of train approaches in a railway network at the micro level

Specifically, each track circuit group within the train route is used to calculate the process of train occupation of track resources, while each individual track circuit in the train route is utilized to calculate the train's running time within the route.

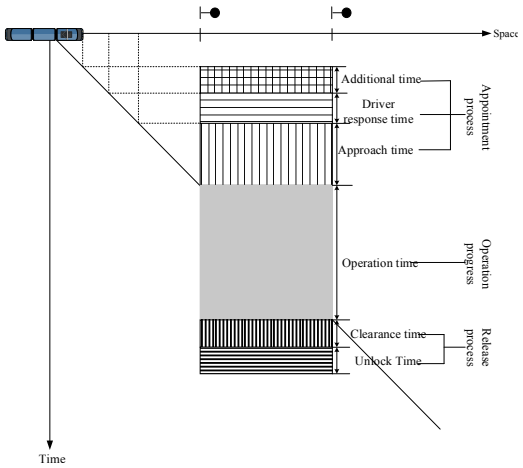


Figure 3: Schematic diagram of the locking time of the track section

The train route blocking time consists of three components: reservation time, running time, and release time. The reservation time includes the time required for setting up the signals and routes before the train enters the route, the driver's observation of the signal lights and reaction time, and the approach time from the indication signal to the entrance signal of the train route. The running time is the period from when the train's front end begins to enter the route (referred to as the train entry time) until the train's front end reaches the end of the route (referred to as the train exit time). The running time is calculated by summing up the running times on the track circuits that belong to the train route. It is assumed that the train can change direction by directly swapping the head and tail on the turnaround route, and thus the running time on the turnaround route can also be calculated based on this principle. The release time is the sum of the clearance time for the train's length and the route unlocking time. The minimum running time of the train on a track circuit is determined based on the length of the track circuit.

Table 1: Traction definition

Symbol	Definition
i, i', j, j'	Station index, $i, i', j, j' \in A$
(i, i')	Interval index, $(i, i') \in G$
l	Train scheme line index, $l \in L$
k, k'	Train index, $k, k' \in K$
t, t', t''	Discrete time unit index in the main space-time network, $t, t', t'' \in T'$
τ, τ', τ''	Extend discrete time unit index in spatiotemporal network, $\tau, \tau', \tau'' \in T''$
(i,t)	The index of the spatiotemporal nodes in the primary spatiotemporal network, $(i, t) \in V$
(i, i', t, t')	The spatiotemporal arc index in the main spatiotemporal network, $(i, i', t, t') \in E$
(i, τ)	Extend the spatiotemporal node index in the spatiotemporal network, $(i, \tau) \in V'$
(i, i', τ, τ')	Extend the spatiotemporal arc index in the spatiotemporal network, $(i, i', \tau, \tau') \in E'$

3.2 Discrete Space-Time Network Approach

To model the periodic train timetabling problem, the train operation plan for the section A-B-C-D over a time span of $H \cdot T$ is represented in a discrete space-time network. In this model, the movement and dwell of trains between stations are depicted as directed arrows (directed edges), as illustrated in 4 and 5, with the unit time length set to 1 minute. Here, A, B, C and D denote station names. Intermediate stations B and C are virtualized into two stations: B', B'' and C', C''. The first virtual station (B' or C') represents the 'arrival' at the station, while the second virtual station (B'' or C'') represents 'departure' from the station. If a train does not stop at the station, it directly proceeds to the second virtual station and departs from there, as exemplified by Train 1 passing through Station B. In contrast, if the train stops at the station, it first arrives at the first virtual station to wait and then departs from the second virtual station, as demonstrated by Train 1 passing through Station C.

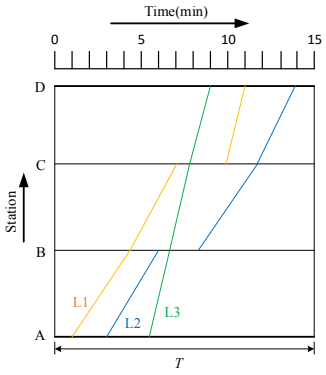


Figure 4: Regular periodic train timetable

Table 2: Collection definitions

Symbol	Definition
A	Station collection, including virtual stations
G	A collection of compartments, including virtual compartments
L	Train scheme line set
K	Train set
K_l	A collection of trains belonging to train scheme line l
V	A collection of space-time nodes in the main space-time network
V'	Extend the collection of spatiotemporal nodes in a spatiotemporal network
V_k	Extend the collection of spatiotemporal nodes in a spatiotemporal network
V'_k	A possible set of space-time nodes in an extended space-time network for train k
E	A collection of space-time arcs in the main space-time network
E'	Extend the collection of space-time arcs in a space-time network
E_k	The set of space-time arcs of train k in the main space-time network
E'_k	The set of spatiotemporal arcs of train k in an extended space-time network
T	A collection of discrete time units in a period, and the size of the set is the length of the period
T'	A collection of discrete time units in a primary space-time network
T''	Extend the collection of discrete time units in a spatiotemporal network
H	Periodic transformation coefficient of T
H'	Periodic transformation coefficient of T'
H''	Periodic transformation coefficient of T''
$\phi(j, j', \tau'')$	Extend the set of spatiotemporal arcs in the spatiotemporal network that are incompatible between interval (j, j') and time τ''

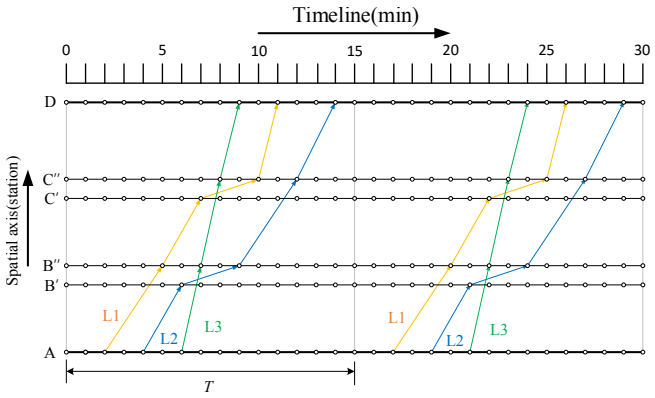


Figure 5: Periodic train diagram in a discrete space-time network

When modeling the train timetable optimization problem using the discrete space-time network approach, let V and E represent the sets of space-time nodes and space-time arcs in the discrete space-time network, respectively. Additionally, V_k and E_k denote the sets of space-time nodes and space-time arcs that train k may occupy. Furthermore, if $t, t', t'' \in T$ are used to index the discrete time units within the planning horizon T, then $(i, t) \in V$ and $(i, i', t, t') \in E'$ are used to index the sets V and E, respectively. Meanwhile, the set $\phi(j, j', t'')$ represents the set of space-time arcs in E that are mutually incompatible on section (j, j') at time t''.

$$\min Z = \sum_{k \in K} \sum_{(i,i',t,t') \in E_k} c_k(i, i', t, t') \cdot x_k(i, i', t, t') \quad (1)$$

$$\sum_{i,t:(i,i',t,t') \in E_k} x_k(i, i', t, t') - \sum_{i,t:(i',i,t',t) \in E_k} x_k(i', i, t', t) = \begin{cases} -1 & i' = o_k, t' = \text{dep}_k^s \\ 1 & i' = d_k, t' = T \\ 0 & \text{otherwise} \end{cases}, \quad \forall l \in L$$

$$\sum_{k \in K} \sum_{(i,i',t,t') \in \phi(j,j',t'')} x_k(i, i', t, t') \le 1, \ \forall (j, j') \in G, t'' \in T \quad (2)$$

$$x_k(i, i', t, t') \in \{0, 1\}, \ \forall k \in K, (i, i', t, t') \in E \quad (3)$$

Equations 1 to 3 present a 0-1 integer programming model for optimizing the macroscopic non-periodic train timetable based on the discrete space-time network. The form of this model, apart from the objective function, is similar to that of the train timetable optimization model proposed by Caprara. The objective function in Equation 1 aims to minimize the total train operating cost, which can encompass various optimization objectives such as train travel time and energy consumption. In this section, the cost of using a space-time arc is set as the corresponding train's running time on that arc, making the objective function the minimization of total train travel time. Constraint corresponds to the flow balance relationship, ensuring that each train selects a unique space-time path. Specifically, Constraint uses the feasible space-time arc set E_k for train k, rather than the set E containing all space-time arcs. Therefore, by defining the set of space-time arcs that train k may traverse, the number of space-time arcs that need to be searched can be reduced, and unnecessary space-time arcs can be eliminated for each train k to meet dwell operation requirements. Constraint is the track capacity constraint, which

Table 3: Parameter definitions

Symbol	Definition
o_k	The station from which train *k* departs
d_k	The final station of train *k*
h_{dd}	The safety interval between two trains departing from the same station in the same direction, and the safety interval time parameters for the rest of the trains. Including $h_{aa}, h_{ap}, h_{pp}, h_{pd}, h_{pa}$
$c_k\,(i, i', t, t')$	The cost of using space-time arc (i, i', t, t') for train *k*
m_l	The frequency of the train scheme line *l*
w_l	The first train with the earliest departure time is included in the train plan line
Q	The number of epochs in the main space-time network, which has $T' = Q \cdot T$ and $T'' = 2Q \cdot T$ for the given parameter *Q*
ϑ	The integer parameter is used to index each master plan in the expansion plan. $\vartheta \in \{0, \cdots, Q\}$
$q_{l,k}$	An integer parameter is used to specify the order in which the train *k* is in the train scheme line *l*. $q_{l,k} \in \{0, \cdots, m_l - 1\}$

Table 4: Variable definitions

Symbol	Definition
$x_k\,(i, i', t, t')$	0-1 Main plan space-time arc selection variable, if train k select space-time arc (i, i', t, t') , $x_k\,(i, i', t, t') = 1$, or $x_k(i, i', t, t') = 0$
$y_k\,(i, i', \tau, \tau')$	0-1 Extend plan space-time arc selection variable, if train k select space-time arc (i, i', τ, τ') , $y_k\,(i, i', \tau, \tau') = 1$, or $y_k(i, i', \tau, \tau') = 0$

describes the space-time resource occupancy constraints in the railway network based on the set train safety headway times. Specifically, only one train can occupy a space-time arc in the set $\phi(j, j', t'')$. Finally, Constraint specifies the type of space-time arc selection variables.

The set *A* denotes the collection of stations in a high-speed railway network, while the set *G* comprises all sections connecting pairs of adjacent stations. The periodic railway timetabling problem involves scheduling a set of train service lines $l \in L$ to repeat periodically with a period length *T*. For each train service line *l*, its operating frequency m_l within the period *T*, stopping pattern, and operational range are predefined. The frequency m_l requires that m_l identical trains be operated within period *T*, distributed uniformly at equal intervals. This uniformity constraint is termed the regularity requirement of periodic timetables.

Specifically, the time interval between the arrival or departure times of any two trains belonging to the same service line at identical stations must be an integer multiple of $\lfloor T/m_l \rfloor$. Zhang explored relaxing the regularity requirement to enhance scheduling flexibility from the perspective of railway line capacity analysis. However, this chapter strictly enforces the regularity requirement to facilitate subsequent modeling and solution procedures.

Additionally, the minimum and maximum running times of trains in sections and dwell times at stations are prescribed. The starting and stopping additional times of trains can be incorporated into the minimum and maximum section running times, provided the stopping patterns are predefined. However, accounting for these constraints may result in significant disparities in actual running times between trains traversing the same section. To prevent overtaking of slower trains by faster ones within a single section, this study adopts the approach of Lie, which involves splitting train running arcs in long sections by inserting a virtual station at the midpoint where overtaking might occur. Consequently, the starting and stopping additional times must be integrated into the minimum and maximum running times for the first and second virtual subsections, respectively.

In this model, block sections are treated as individual track circuits, each consisting of two nodes numbered sequentially from left to right (for clarity, this numbering does not follow the conventional switch numbering rules). An example is provided in 6.

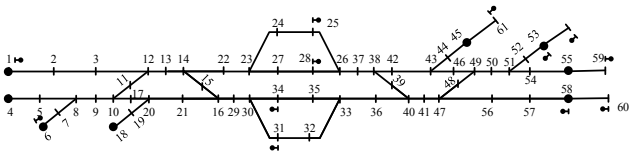


Figure 6: Numbered refined railway lines

The periodic train timetable problem essentially involves periodically scheduling a series of planned train space-time paths within each period of length *T*. Each path is associated with the operating frequency, stopping patterns, and running zones of the train. Additionally, to meet the periodic regularity requirements of trains, the arrival or departure times of any two trains with parallel paths in the space-time network must maintain a fixed time interval. Furthermore, the running time of trains on each track circuit (including dwell time if the train stops) is constrained within a specific time range. Under a given stopping pattern, trains can precisely control their acceleration and deceleration to manage time within the minimum

and maximum running time limits. The railway line N is represented as a microscopic discrete space-time network graph as follows: the vertex set V consists of individual track circuits (or track circuit groups) in the railway network, and the directed arc set C represents the running lines through each track circuit. A railway operations department needs to schedule train timetables for three different physical routes. Since this study considers a microscopic-level railway network, both stations and sections can be represented by track circuits, which are further composed of nodes. Thus, the network can represent the entire railway system, with edge nodes representing train origin and destination nodes. Physical Route 1 is depicted by a blue dashed line, where the train departs from Node 1, briefly stops at track circuit (24, 25), and finally arrives at track circuit (45, 61) via Node 45. Physical Route 2 and Physical Route 3, represented by orange and red dashed lines, respectively, depart from Node 6 and Node 1, pass through the track circuits without stopping, and ultimately arrive at track circuit (56, 59) via Node 56, as shown in 7.

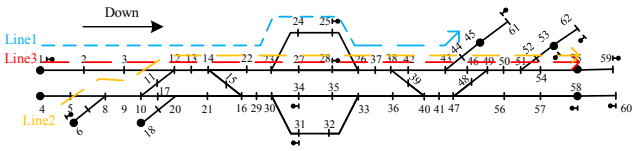


Figure 7: A railway line with edge nodes that contains three paths

In this model, the passage of a train through each track circuit involves three processes of the track section locking time: reservation process, running process, and release process. 8 illustrates the scenario where two consecutive trains from different service lines pass through two track circuits. In the space-time network, the directed arcs are connected end-to-end. Taking the black service line as an example, the two directed arcs represent the running process time of the train passing through track circuits, which is the time difference from when the train head enters the starting node of the track circuit to when the train tail leaves the ending node of the track circuit. This includes both the pure running time and the dwell time of the train.

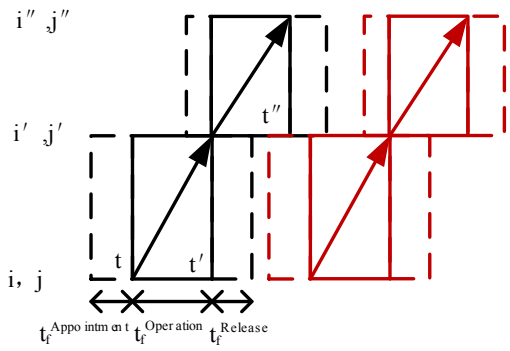


Figure 8: Directed arcs in discrete space-time networks

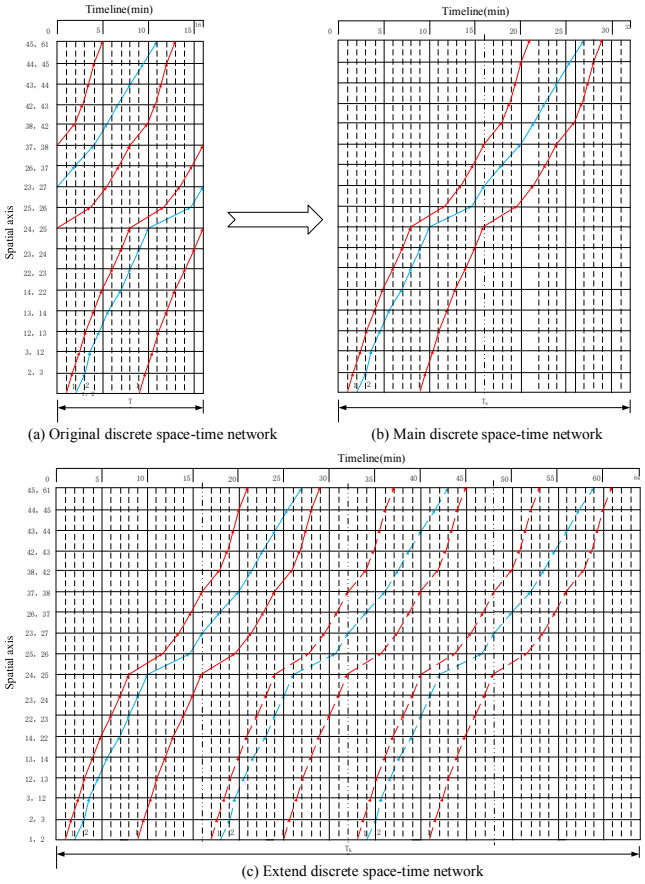


Figure 9: Three discrete space-time networks

3.3 Model Assumptions

To facilitate processing and comprehension, several assumptions are made based on the scope and nature of the problem in this study:

1. The minimum time granularity of the space-time network model in this project is assumed to be 15 seconds. The time axis is divided into unit-length multiples, and a smaller minimum time granularity can be applied for higher precision.

2. The railway line in this problem is double-tracked. The model considers only trains in the down direction, assuming no conflicts with trains in the up direction.

3. All trains are assumed to depart from the initial peripheral nodes of the railway network and terminate at designated peripheral nodes.

4. Trains within the same service line are assumed to follow identical physical paths (i.e., traverse the same track circuit sections), with their space-time paths uniformly distributed within the period.

3.4 Periodic Timetabling Optimization Model

9 illustrates a periodic timetable with two distinct train service lines, both traversing Physical Path 1 on the railway line, thereby sharing the same track circuits by default. The period length T is 16 minutes, where Service Line 1 corresponds

to a train frequency of 2 (i.e., two trains depart within one period), and Service Line 2 corresponds to a frequency of 1. Additionally, trains make stops at track circuits (24, 25), with dwell time included in the total travel time. Within a single period, the temporal spans of the space-time paths for all three trains exceed the period boundary. Consequently, the original discrete space-time network cannot display the complete continuous service lines. Instead, portions of the space-time paths outside the current period are mapped into the period's space-time network, as shown in 9(a).

Next, we transform the original discrete space-time network into a master discrete space-time network (hereafter referred to as the 'master graph') to visualize the full continuous service lines. This is achieved by extending the timetable's temporal horizon to $H \cdot T$, forming a master space-time network with an expanded temporal span. Within this master network, space-time arcs of the same path across multiple periods are concatenated, retaining only one complete service line, as depicted in 9(b). The parameter *H* is set to the smallest integer satisfying $H \cdot T > T_{max}$, where T_{max} denotes the maximum possible travel time for any train to reach its destination. For example, in 9(b), T_{max}= 29 minutes. Thus, $H = [29/16] = 2$, resulting in a master temporal span T'= 32 minutes.

Finally, to reflect the periodicity of train service lines and ensure conflict-free periodic timetables, we further extend the temporal axis of the master graph to generate an extended discrete space-time network (hereafter the 'extended graph'). Assuming the extended period length is $T'' = H' \cdot T$, where $H' \geq H$, the extended graph replicates each service line H' times, with each replication shifted by *T* units along the temporal axis. As shown in 9(c), when H'= 2 and T''= 64 minutes, the service lines from the master graph are replicated twice, represented by dashed space-time arcs. In the extended graph, although space-time paths may cross the right boundary of the first period, the departure times of the first trains in all service lines must lie within the interval [0,*T*). Furthermore, the departure time window Δt for the first train in Service Line *l* is constrained to $\Delta t \subseteq [0, T/m_l)$, where m_l is the service frequency. According to the Lemma, subsequent trains in the same service line are uniformly distributed with equal intervals T/m_l, forming parallel directed arcs in the space-time network. For example, the first and second trains on Path 1 depart at 1 minute and 9 minute, respectively, with an interval of 16/2=8 minutes.

3.5 Model constraint

$$\min Z_1 = \sum_{k \in K} \sum_{(i,i',\tau,\tau') \in E'_k} c_k\left(i,i',\tau,\tau'\right) \cdot y_k\left(i,i',\tau,\tau'\right) \quad (4)$$

$$\sum_{i,t:(i,i',t,t') \in E_{w_l}} x_{w_l}\left(i,i',t,t'\right) - \sum_{i,t:(i',i,t',t) \in E_{w_l}} x_{w_l}\left(i',i,t',t\right) = \begin{cases} -1 & i' = o_{w_l}, t' = dep_k^s \\ 1 & i' = d_{w_l}, t' = T' \\ 0 & \text{otherwise} \end{cases}, \quad \forall l \in L$$

$$y_k\left(i,i',\tau,\tau'\right) = x_k\left(i,i',t+\vartheta T, t'+\vartheta T\right) \ \forall k \in K, (i,i',t,t') \in E, (i,i',\tau,\tau') \in E', \vartheta \in \{0,\cdots,Q\}, \tau = t+\vartheta T, \quad (5)$$

$$\tau' = t' + \vartheta T \quad (6)$$

$$\sum_{k \in K} \sum_{(i,i',\tau,\tau') \in \phi(j,j',\tau'')} y_k\left(i,i',\tau,\tau'\right) \leq 1, \quad \forall \left(j,j'\right) \in G, \tau'' \in T''$$

$$x_k\left(i,i',t,t'\right) \in \{0,1\}, \ \forall k \in K, (i,i',t,t') \in E \quad (7)$$

$$y_k\left(i,i',t,t'\right) \in \{0,1\}, \ \forall k \in K, (i,i',t,t') \in E \quad (8)$$

In terms of mathematical representation, space-time arc selection variables $x_k(i,i',t,t')$ and $y_k(i,i',t,t')$ are designed for the master plan and the extended plan, respectively. Specifically, the time units in the master space-time network are indexed by subscripts t and t', while the time units in the extended space-time network are indexed by subscripts τ and τ'. Equations 4 to 8 present the periodic train timetable optimization model based on the extended space-time network. In the model, the objective function 4 minimizes the total travel time of all trains in the extended plan. Constraint is the flow balance constraint, ensuring that the first train w_l in each train service line $l \in L$ can find a unique space-time path in the master space-time network. The variable $x_{w_l}(i,i',t,t')$ in constraint represents the space-time arc selection variable for the first train w_l in the master space-time network. Constraint generates the space-time paths for the remaining trains in the same train service line $l \in L$, excluding the first train w_l. These paths are obtained by shifting the space-time path of the first train w_l by integer multiples of the interval $min\left\{\lfloor T/m_l \rfloor, T-1\right\}$. Additionally, the integer parameter $q_{l,k}$, in constraint specifies the order of train *k* in train service line *l*, where $q_{l,k}$, belongs to the set $\{1,\cdots,m_l-1\}$, thereby excluding the first train w_l from constraint.

Another key contribution of the periodic train timetable optimization reconstruction model based on the extended space-time network is the use of variable separation and replication techniques. Constraint 6 is the consistency constraint between the master plan and the extended plan, which replicates the train space-time paths in the master plan *Q*+1 times to form the extended plan. Specifically, the integer parameter ϑ in constraint 6 controls the number of copies of the master plan required to form the extended plan, where ϑ belongs to the set $\{0,\cdots,Q\}$. The form of constraint 6 is similar to the nonanticipativity constraint in two-stage stochastic mixed-integer programming models that link the first and second-stage decisions. Crainic have demonstrated that the progressive hedging approach can effectively handle such constraints in two-stage stochastic mixed-integer programming models. Furthermore, since constraint 6 in the model only performs replication operations, there is no need to relax its dual into the objective function. Constraint **??** is the track capacity constraint, ensuring no train conflicts occur within the planning horizon of the extended space-time network, thereby guaranteeing the feasibility of the original periodic train timetable. Constraints 7 and

8 define the types of space-time arc selection variables in the master plan and the extended plan.

In addition to the aforementioned constraints, the representation of train timetables in discrete space-time networks inherently implies several fundamental constraints, including:
(1) Adjacent Track Circuit Arrival-Departure Temporal Constraint: The sequential connection of space-time arcs for the same train service ensures temporal continuity between adjacent track circuits. Specifically, the departure time from the preceding track circuit equals the arrival time at the subsequent track circuit.
(2) Space-Time Network and Physical Path Correspondence Constraint: The existence of a directed space-time arc in the discrete network explicitly indicates that the train selects track circuit (*i, j*) along its route from the origin to the destination node in the railway network, while the absence of such an arc implies non-selection.
(3) Running and Dwell Time Constraint: The traversal of track circuit (*i, j*) inherently incorporates both running time and dwell time. By default, the total time expenditure on the directed arc equals the temporal difference between the arrival time at the subsequent track circuit and the arrival time at the current track circuit. Consequently, these constraints are not explicitly formulated in the model construction.

4 Solution algorithm

4.1 Grouping by Path Similarity

The optimization of detailed operational timetable compilation requires modeling the railway network at a micro level. Although this approach allows for more efficient utilization of track conditions, it also significantly increases the complexity of the problem. To address this issue, reduce the complexity, and accelerate the algorithmic solution process, this paper proposes to establish methods for grouping trains.

Herrigel, while grouping passenger trains of the Swiss railway to solve the PESP model, proposed the geographical grouping (Geo) method based on the practical scheduling requirement of assigning railway staff to different regions to organize trains. Drawing on this idea, in the microscopic-level railway network, since the nodes included in the physical paths of different trains are not identical, we can group trains based on the similarity of the node sets they pass through. Cluster analysis, also known as classification analysis, is a statistical method that divides original objects into multiple relatively homogeneous groups, with the classes unknown prior to the analysis. Clustering partitions samples into several groups based on their distances or similarities, aiming to minimize intra-group distances while maximizing inter-group distances. Among clustering algorithms, K-Means is one of the most widely used. Since the input data is unlabeled, it belongs to unsupervised learning.

$$S\left(l_i, l_j\right) = 1 - \frac{N_{l_i \cap l_j}}{\left(N_{l_i} + N_{l_j}\right)/2}, \forall l \in L \tag{9}$$

Where the similarity distance $S \in [0,1]^{L \times L}$, L is a collection of all train routes, N_{l_i}, N_{l_j} are the number of nodes that path l_i, l_j passed.

Setting *k=m* involves dividing all trains into mm groups. A heuristic algorithm is designed to cluster and group the train paths, with the principle that the resulting groups should have approximately equal numbers of elements, and the elements within each group should minimize the total similarity with other elements in the same group. In other words, the most similar paths are assigned to the same group based on the similarity measure. The flowchart of the algorithm is shown in 10. The steps of this heuristic algorithm are as follows:
Definitions:

- Sample Set: $L = \{l_1, l_2, \ldots, l_n\}$
- Cluster Groups: $K = \{K_1, K_2, \ldots, K_m\}$

Procedure:
1.Initialization:

All cluster groups $K_j\,(j = 1, 2, \ldots, m)$ are initialized as empty sets.
2. Seed Selection:

- Select two elements $l_a, l_b \in L$ with the maximum pairwise dissimilarity (i.e., $\mathrm{argmax}_{l_i, l_j} d\left(l_i, l_j\right)$,where d denotes the distance metric).

- Assign l_a and l_b as initial centroids to distinct clusters K_1 and K_2 .

- Iteratively select the remaining $m-2$ elements from L to maximize the minimum inter-cluster dissimilarity,ensuring the sum of pairs similarities S is minimized.

Formally:

Select $l_k \in L \setminus \{K_1 \cup K_2\}$ such that $\sum_{l_i \in K_j} S\left(l_k, l_i\right)$ is minimized for all j .

Assign these elements to new clusters $K_3, \ldots, K_m$.
3. Balanced Assignment:

- While $L \neq \varnothing$:

a. Identify the cluster $K_{\min}$ with the smallest cardinality.

b. Select an element $l \in L$ that minimizes the intra-cluster similarity when added to $K_{\min}$: $l^* = \mathrm{argmin}_{l \in L} \sum_{l_i \in K_{\min}} S\left(l, l_i\right)$.

c. Assign l^* to $K_{\min}$ and remove l^* from L .

4.2 Sorting by Occupation Time

Since different groups occupy the railway line for varying total durations, prioritizing the solution of groups with longer occupation times can be beneficial. This is because their longer spans reduce the scale of the directed arc options in the space-time network for subsequent groups, thereby narrowing the search range during computational solving and accelerating the solution process.

The occupation time of a train group is the sum of the occupation times of all trains within the group, where the occupation time of a single train is the total time it occupies all track circuit sections it passes through, including both running and dwell times. Although the occupation times of different trains may overlap temporally, since the timetable is not yet

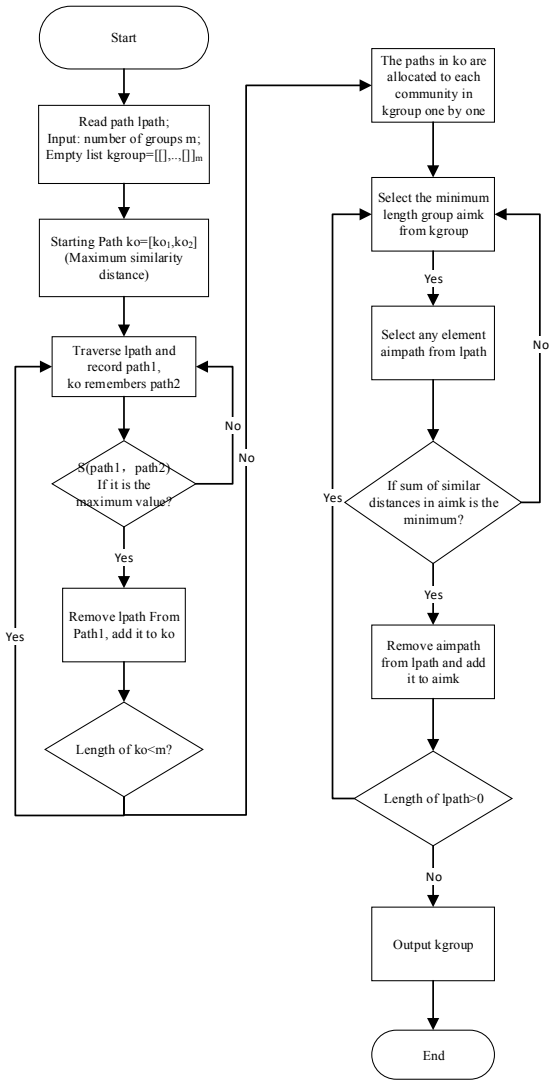


Figure 10: Flow diagram of path similarity grouping algorithm

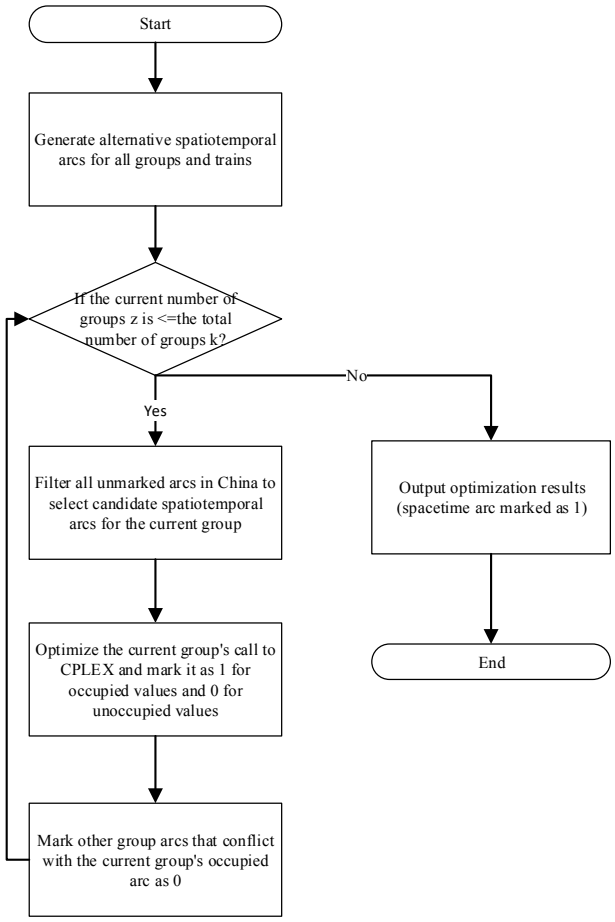


Figure 11: Flow diagram of the grouped iterative solution algorithm

determined when using this sorting method, the sum of occupation times is used as the sorting criterion. The formula for calculating the occupation time of a train group is as follows:

$$t_i = \sum_{a \in A_i} \sum_{link \in path_a} \left(\frac{L_{link}}{\alpha \cdot v_m^{link}} + t_{dw} \right) \tag{10}$$

Where the length of the track circuit segment L_{link} , the maximum allowable speed at which a train can pass through a section of the track circuit is v_m^{link} , the train stop times is t_{dw}, the velocity coefficient is α.

The flowchart of the algorithm is shown in 11. After completing the train grouping and sorting, it is essential to design a method for iteratively incorporating the grouped trains into the model for solving. After each target group optimization, the determined arcs are added to the ‘Marked 1’ list. During the next optimization, the arcs from the latest ‘Marked 1’ list, along with all arcs of the group to be optimized, are treated as the target optimization arcs. Constraints are added to the model to ensure that all arcs in the ‘Marked 1’ list are fixed to 1. After optimization, this list is updated accordingly.

5 Case study

This chapter adopts the network data from the 2016 RAS Problem Solving Competition (Railway Application Section of the Institute for Operations Research and the Management Sciences (INFORMS)) titled 'Train Scheduling in Railway Networks: Integrated Optimization of Timetabling and Maintenance Task Allocation'. By configuring rational parameters and leveraging the mathematical models established earlier, this study solves the network to generate a refined periodic railway timetable for practical operations. The grouping algorithm, model-solving procedures, and timetable generation in this work are implemented using Python 3.7.12.

5.1 Case description and parameter configuration

The railway network provided in the 2016 RAS competition comprises 27 stations, 55 track segments, 261 routes, 1,027 track circuit section groups, 1,811 track circuit sections, and 1,619 nodes. The network is partitioned into five geographical divisions: Western, Eastern, Northern, Southern, and M-Station. The M-Station area represents the most complex hub, consisting of 19 siding tracks and 4 main tracks, interconnected with all four regional divisions. Consequently, most trains in the network traverse M-Station. The network also includes maintenance track circuit section groups, located in stations or track segments.

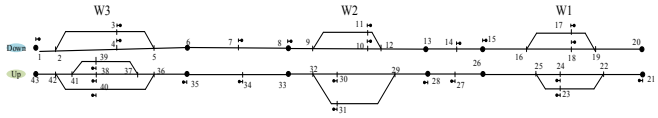


Figure 12: Two-track rail route map of the western region

Due to the excessively large number of track circuit sections in the entire railway network, the generated detailed timetable is inconvenient to display. This example extracts the western region of the network, focusing on three stations as the target line instance. To reduce the number of track circuit sections along the line, we combine several track circuit sections, as shown in 12. This regional line is a double-track railway comprising 43 nodes, 48 track circuit sections, and 3 stations (W1-W3). Since the model does not consider conflicts between trains in different directions, only the track circuit sections that may be passed by trains traveling in the downward direction from W3 to W1 are retained, totaling 14 sections. Additionally, as the mathematical model focuses solely on optimizing the train timetable, the original maintenance sections are not considered in this project.

The input data files include information on nodes, track circuit sections, track circuit section groups, and train details, with their specific meanings described in 5.

5.2 Parameter configuration

(1) Fixed Parameters

Following the methodology of Meng for setting occupation and release intervals of railway resources, the reservation time for track circuit sections is set to 60 seconds, meaning a track circuit section is occupied one minute before a train arrives. Similarly, the release time is set to 60 seconds, indicating the track circuit section remains occupied for one minute after the train departs.

In this case study, to avoid excessive computational complexity from a large number of space-time arcs, the dwell time at stations is fixed at 120 seconds, representing the minimum allowable dwell time. Additionally, the minimum traversal time for a train to pass a track circuit section is calculated based on the section length, maximum permitted speed, and train speed coefficient. The maximum traversal time is defined as 120% of the minimum traversal time, forming a candidate set of travel times for each train across track circuit sections. This approach effectively reduces the number of space-time arcs to be solved, particularly in networks with long or densely segmented tracks. The base period length of the periodic timetable is set to 2 hours (7200 seconds). Notably, the minimum time granularity is defined as 10 seconds, requiring all time-related variables in the case study to be converted into unit-scale quantities (e.g., the period duration is represented as 720 units). Finally, the master plan space-time network is extended to generate an extended plan space-time network, reflecting the regularity of train service lines. A minimum scaling factor of 2 is applied, fixing the extended period length as twice the master plan period length.

(2) Variable Parameters

After defining all fixed parameters, the period scaling factor (from the base plan to the master plan) is adjusted to finalize the periodic timetable configuration. The master plan period length is determined by ensuring that the latest arrival time of any service line (with its first train departing within the initial period) does not exceed the master plan period length.

The case study involves 8 trains, which can be grouped into 2, 3, or 4 clusters. The grouping principle prioritizes minimizing total computational time while ensuring that no single cluster contains an excessive number of trains, which would negate the purpose of the proposed grouping optimization framework.

5.3 Validation of train grouping methodology

To visually compare the performance of direct optimization versus group-based optimization, we designed two train path scales: a small-scale group (hereafter 'Group 1') with an average path length of 5 track circuit sections per train, and a large-scale group (hereafter 'Group 2') with an average path length of 10 track circuit sections per train. The selected track circuit regions for these groups are highlighted by the blue and red boxes in 13. Notably, each additional track circuit section doubles the number of space-time arcs generated per train. For example, a single train in Group 2 generates 31 times more space-time arcs than one in Group 1. For 6 trains, this difference escalates to 186 times.

The total number of trains in both groups was varied dynamically from 2 to 6. Using identical hardware, direct optimiza-

Table 5: The road network data for the case

Filename	Attribute	Meaning
Input_Node	node_id	Node number
Input_Link	link_id	Track circuit number
	from_node	Track circuit start node
	to_node	Track circuit end node
	length_in_mile	Track Circuit Line Length (miles)
	speed_limit_in_mph_FT	Pass the maximum speed limit (downlink)
	dwelling_allowable_flag	Whether trains are allowed to stop
Input_Cell	cell_id	The track circuit group number
	including_link	Included track circuits
Input_Train_Info	train_id	The train number
	origin_node_id	The starting point of the train
	destination_node_id	End of the train
	speed_multiplier	Velocity coefficient
	frequency	The frequency of the train during the original period
	link_of_ actual_path	The track circuit through which the physical path of the train passes in turn
	dwelling_link	Trains need to stop at the track circuit in turn

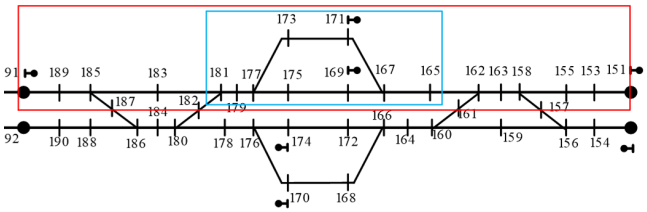


Figure 13: Track circuit section in the experimental area

tion and group-based optimization were applied to each experimental group. For group-based optimization, trains were randomly grouped, with a maximum of 2 trains per group to ensure computational tractability. The total computation times for model optimization in both experimental groups are compared in 14 and 15.

Key findings from the experiments include:

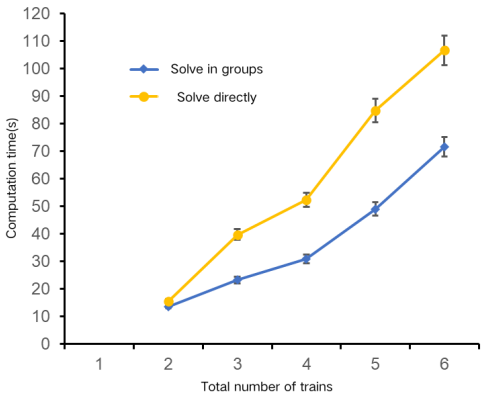


Figure 14: The first group of experiments computation time

1. Group-based optimization consistently reduced computation time compared to direct optimization, regardless of path scale.

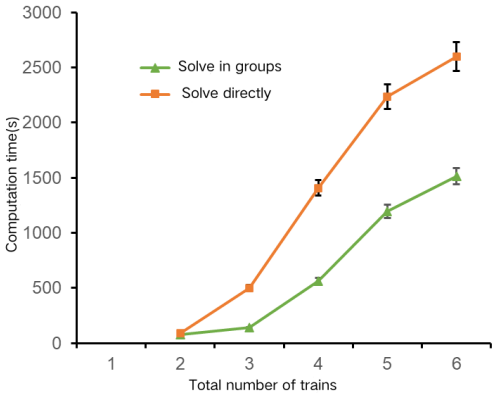


Figure 15: The second group of experiments computation time

2. As the total number of trains increased, computation time rose for both methods. However, the advantage of group-based optimization became more pronounced with larger fleets. For 6 trains, direct optimization exhibited significantly longer computation times, with the gap widening progressively.

3. For the same total number of trains, longer train paths (Group 2) amplified the time saving benefits of group-based optimization

4. Objective function values from both methods were nearly identical, confirming that group-based optimization preserves solution quality.

In summary, the results demonstrate that group-based optimization significantly reduces computation time for larger fleets, highlighting its practical superiority. Furthermore, the method exhibits even greater advantages when applied to large-scale networks with numerous track circuit sections and high train volumes.

5.4 Experimental design and results

To evaluate the effectiveness of different train grouping methods and group sequencing strategies, we conducted comparative experiments. The experimental case uses the track circuit sections within the blue-boxed region of 13, with a total of 12 trains.

(1) Comparison of Train Grouping Methods

Three grouping strategies were tested:

1. Random Grouping: Trains are randomly assigned to groups, with a maximum of 3 trains per group.

2. Path Similarity-Based Grouping: Trains sharing overlapping track circuit sections or similar physical paths are grouped together.

3. Departure-Time Clustering-Based Grouping: Trains are clustered using the K-means algorithm based on their scheduled departure times.

A total of 12 trains (representing 12 distinct service lines) were divided into 2 to 6 groups under three grouping strategies: route similarity-based grouping, frequency-based grouping, and random grouping. To control for extraneous variables, the sequencing of train groups was uniformly randomized across all configurations. Each configuration underwent three repeated trials, resulting in 45 total experiments. The objective function value remained consistent at 44,490 s in all trials. The averaged computation times are summarized in 6 and visualized in 16. The final result reveals that the route similarity-based grouping significantly outperformed both frequency-based grouping and random grouping in terms of average computation time, while the latter two methods exhibited comparable performance. All grouping methods preserved solution quality, without impacting the objective function value. Across all grouping strategies, average computation time increased proportionally with the number of groups, highlighting a trade-off between granularity and computational efficiency.

(2) Comparison of Train Group Sequencing Methods

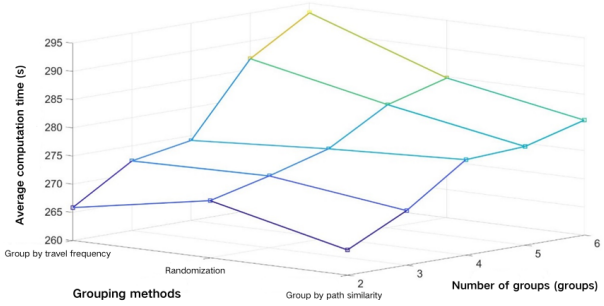


Figure 16: Chart of the results of the grouping method control experiment

The train grouping method of controlling irrelevant variables is divided into 6 groups by the method of route similarity. After the grouping is completed, the trainsets are sorted by random sorting and the total occupancy time. The total occupancy time of each group was calculated according to equation 10, and the experiment was repeated 3 times, and the sorting results and average calculation time are shown in 7.

The results show that the average calculation time is nearly 1.233% compared with random sorting, and it does not affect the objective function value, which proves that sorting by occupancy time from large to small can improve the computational efficiency of the model to a certain extent.

5.5 Solution of the practical case

In the example, the total length of the line is approximately 65 km. To simplify the calculation, only trains traveling in the downward direction are considered. As shown in 17, based on whether the trains stop at intermediate stations, there are two train operation paths (via nodes):

(1) With stops: $3 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 11 \rightarrow 12 \rightarrow 13 \rightarrow 14 \rightarrow 15 \rightarrow 16 \rightarrow 17$

(2) Without stops: $3 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10 \rightarrow 12 \rightarrow 13 \rightarrow 14 \rightarrow 15 \rightarrow 16 \rightarrow 17$

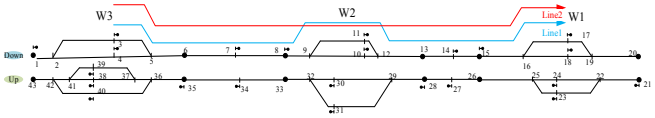


Figure 17: Actual train route map

The trains operate at a maximum speed of 180 km/h and are categorized into four types based on their stopping patterns at intermediate stations and line-specific conditions:

1. Type I: Non-stop at intermediate station W2 with a speed coefficient of 1;

2. Type II: Non-stop at intermediate station W2 with a speed coefficient of 0.7;

3. Type III: Stops at intermediate station W2 with a speed coefficient of 1;

4. Type IV: Stops at intermediate station W2 with a speed coefficient of 0.7.

Key operational characteristics of these train types, including acceleration/deceleration profiles, dwell times, and energy consumption metrics, are summarized in 8. According to the path similarity grouping method, combined with the train operating frequency, all trains are divided into 4 groups. The train grouping and solving order are shown in 9. 10 shows the detailed timetable for each train that generates the first train entering the track circuit.

Using Python's Matplotlib library, a microscopic-level periodic timetable was generated, as illustrated in 18. Finally, following the methodology of Andrea for generating train block time diagrams, we constructed periodic timetables by filling closed rectangular blocks (aligned with diagonal space-time arcs) with track circuit traversal times or track circuit blocking times. The resulting diagrams are shown in 19 and 20, respectively.

Table 6: Comparison table of experimental results of grouping methods

Groups' number \ Method	Path similarity	Travel frequency	Randomization	Average
2	264.493	270.158	265.852	266.834
3	269.686	272.753	272.336	271.592
4	276.930	275.772	274.185	275.629
5	277.488	281.824	286.893	282.068
6	280.373	284.768	293.256	286.132
Average	273.794	277.055	278.504	–

Table 7: Experiment of trainset sequencing method

	(1) 1,6	(2) 7,12	(3) 2,8	(4) 3,9	(5) 4,10	(6) 5,11	Average calculation time
Sort randomly	3	2	5	6	4	1	280.373
Occupy time sorting	4	3	1	6	2	5	276.917

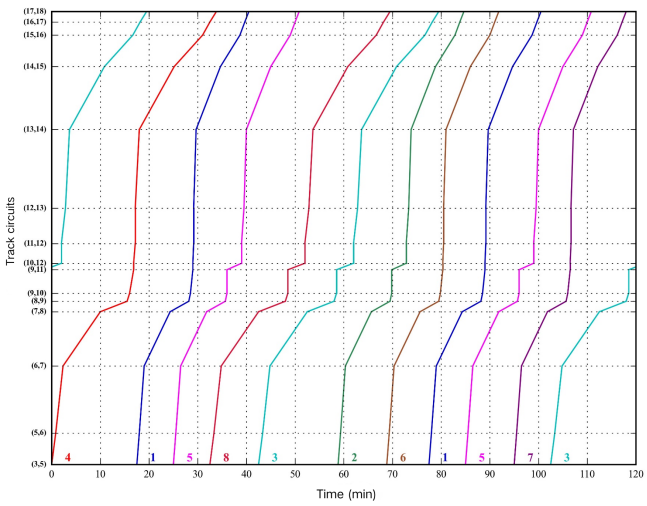


Figure 18: Periodic train timetable

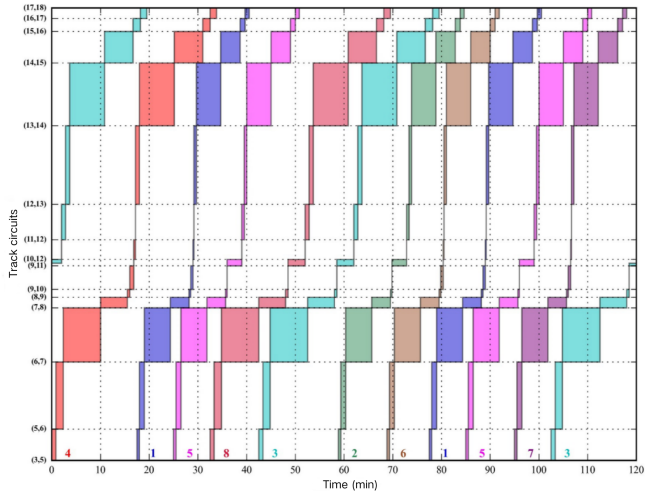


Figure 19: Periodic train timetable (Track circuit section running time)

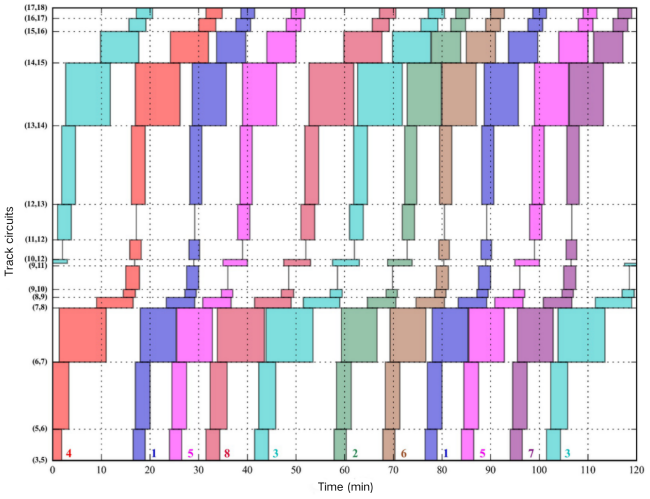


Figure 20: Periodic train timetable (Track circuit section locking time)

6 Conclusion

At the microscopic level, railway stations and track sections are modeled with train movements represented through track circuit units along physical paths. Train occupation of these track circuits is regulated based on locking times to ensure safe operation. By representing time on the horizontal axis and traversed track circuits on the vertical axis, a sequence of space-time network diagrams is constructed: the original, the main, and the extended space-time networks. The mathematical model is formulated using the arcs of the extended space-time network, aiming to minimize total travel time subject to constraints including flow conservation, periodicity coupling, track circuit occupancy, and binary decision variables. Due to its combinatorial nature, the resulting integer programming model is NP-hard. To alleviate computational complexity and enhance solution efficiency, this study proposes various train grouping strategies and ordering heuristics, alongside a train-group iterative solution algorithm. The model is implemented

Table 8: Train information

The train number	Velocity coefficient	Frequency	Pass through the track circuit section	Stop track circuit section
1	1	2	1;2;3;4;5;6;8;10;11;12;13;14	
2	1	1	1;2;3;4;5;7;9;10;11;12;13;14	7
3	0.7	2	1;2;3;4;5;7;9;10;11;12;13;14	7
4	0.7	1	1;2;3;4;5;6;8;10;11;12;13;14	
5	1	2	1;2;3;4;5;7;9;10;11;12;13;14	7
6	1	1	1;2;3;4;5;6;8;10;11;12;13;14	
7	1	1	1;2;3;4;5;6;8;10;11;12;13;14	
8	0.7	1	1;2;3;4;5;7;9;10;11;12;13;14	7

Table 9: Train group information

Group number	The number of the train in the group	Solve order
1	1,6	3
2	2,8	2
3	3,5	1
4	4,7	4

in Python with IBM ILOG CPLEX as the solver.

Model validation employs real-world data from the 2016 RAS problem-solving competition. Eight trains are partitioned into four groups using a path similarity clustering method, which are then optimized sequentially based on descending total occupation times. The grouped approach reduces computation time to 2710.894 seconds, yielding a 28.49% efficiency gain compared to ungrouped solving. Within a 2-hour scheduling horizon, the total travel time for down-direction trains reaches 18860 seconds (5.24 hours), demonstrating effective utilization of line capacity and improved timetable quality.

The optimization of train timetables at such a fine-grained microscopic level not only enhances operational efficiency but also directly benefits daily life by improving punctuality and reliability of rail services. This leads to reduced passenger waiting times, smoother transfers, and increased capacity to accommodate growing travel demand. Consequently, the proposed approach contributes to more sustainable and user-friendly public transportation systems, promoting economic development and enhancing the overall quality of urban mobility.

Virtual Stations and Intelligent Decision-Making: Intermediate stations are virtualized into two stations—one for ”arrival” and one for ”departure.” An AI system can intelligently decide on train stop or through-running strategies at these virtual stations based on real-time passenger flow, line conditions, and train priorities.

Train Paths and Predictive Analytics: Train paths include track circuits and switches. For safety, only one train can occupy a path at any given time. AI models can analyze historical data to predict potential conflict points for different trains on specific paths and perform proactive scheduling optimization.

Train Path Blocking Time and Intelligent Optimization: This includes reservation time (for signal and route setting, driver reaction time, and approach time to the entry signal), running time (from the train's front end entering to its front end exiting the path, calculated by summing running times on track circuits), and release time (clearance time for train length plus route unlocking time). An AI system can dynamically adjust these parameters to adapt to varying operational conditions, such as adverse weather or equipment failures, thereby achieving more flexible and efficient scheduling.

References

[1] Brännlund, U., Lindberg, P. O., NOU, A., and Nilsson, J. E. “Railway Timetabling Using Lagrangian Relaxation,” Transportation Science, vol. 32, no. 4, pp. 358-369, 1998.

[2] Caimi, G., Fuchsberger, M., and Laumanns, K. “A multilevel framework for generating train schedules in highly utilised networks,” PUBLIC TRANSPORT -SPRINGER-, 2011.

[3] Caprara, A., Fischetti, M., and Toth, P. “Modeling and Solving the Train Timetabling Problem,” Operations Research, vol. 50, no. 5, pp. 851-861, 2002.

[4] Caprara, A., Monaci, M., Toth, P., and Guida, P. L. “A Lagrangian heuristic algorithm for a real-world train timetabling problem,” Discrete Applied Mathematics, vol. 154, no. 5, pp. 738-753, 2006.

[5] Caprara, A., Fischetti, M., and Toth, P. “Modeling and Solving the Train Timetabling Problem,” Operations Research, vol. 50, no. 5, pp. 851-861, 2002.

[6] Carey, M. “A model and strategy for train pathing with choice of lines, platforms, and routes,” Transportation Research Part B Methodological, vol. 28, no. 5, pp. 333-353, 1994.

[7] Corman, F., D'Ariano, A., Marra, A. D., Pacciarelli, D., and Sama, M. “Integrating train scheduling and delay management in real-time railway traffic control,” Transportation Research Part E Logistics and Transportation Review, vol. 105c, no. sep., pp. 213-239, 2016.

Table 10: Detailed train schedules

Track circuits	Length (km)	Speed limit (km/h)	Entry time (minutes:seconds)							
			1	2	3	4	5	6	7	8
(3,5)	0.3	40	17:30	58:50	42:30	00:00	25:00	68:50	95:00	32:30
(5,6)	1.2	80	18:00	59:20	43:20	00:50	25:30	69:20	95:30	33:20
(6,7)	12.8	150	19:00	60:20	44:50	02:20	26:30	70:20	96:30	34:50
(7,8)	10.35	180	24:20	65:40	52:30	10:00	31:50	75:40	101:50	42:30
(8,9)	0.4	80	28:10	69:30	58:00	15:30	35:40	79:30	105:40	48:00
(9,10)	0.3	40	28:30	——	——	16:00	——	79:50	106:00	——
(9,11)	0.9	60	——	69:50	58:30	——	36:00	——	——	48:30
(10,12)	0.25	100	29:00	——	——	16:50	——	80:20	106:30	——
(11,12)	0.75	100	——	72:50	62:00	——	39:00	——	——	52:00
(12,13)	1.35	180	29:10	73:20	62:50	17:10	39:30	80:30	106:40	52:50
(13,14)	15	180	29:40	73:50	63:40	18:00	40:00	81:00	107:10	53:40
(14,15)	12	180	34:40	78:50	70:50	25:10	45:00	86:00	112:10	60:50
(15,16)	1.2	80	38:40	82:50	76:40	31:00	49:00	90:00	116:10	66:40
(16,17)	0.5	40	39:40	83:50	78:10	32:30	50:00	91:00	117:10	68:10
(17,19)	0.4	60	40:30	84:40	79:30	33:50	50:50	91:50	118:00	69:30

[8] Cordone, R. and Redaelli, F. "Optimizing the demand captured by a railway system with a regular timetable," Transportation Research Part B Methodological, vol. 45, no. 2, pp. 430-446, 2011.

[9] Crainic, T. G., Fu, X., Gendreau, M., Rei, W., and Wallace, S. W. "Progressive hedging-based metaheuristics for stochastic network design," Networks, vol. 58, no. 2, pp. 114-124, 2011.

[10] D'Ariano, A., Pacciarelli, D., and Pranzo, M. "A branch and bound algorithm for scheduling trains in a railway network," European Journal of Operational Research, vol. 183, no. 2, pp. 643-657, 2007.

[11] Dessouky, M. M., Quan, L. U., and Zhao, J. "An exact solution procedure to determine the optimal dispatching times for complex rail networks," Iie Transactions, vol. 38, no. 2, pp. 141-152, 2006.

[12] Dollevoet, T. T., Corman, F., D'Andriano, A., and Huisman, D. "An Iterative Optimization Framework for Delay Management and Train Scheduling," Flexible Services and Manufacturing Journal, vol. 26, no. 4, pp. 490-515, 2012.

[13] Dorfman, M. J. and Medanic, J. "Scheduling trains on a railway network using a discrete event model of railway traffic," Transportation Research Part B Methodological, vol. 38, no. 1, pp. 81-98, 2004.

[14] Kroon, L., Huisman, D., Abbink, E., Fioole, P. J., Ybema, R., Fischetti, M., Maroti, G., Schrijver, A., and Steenbeek, A. "The new Dutch timetable: The OR revolution," Operations Research, 2009.

[15] Gattermann, P., Großmann, P., Nachtigall, K., and Schöbel, A. "Integrating Passengers' Routes in Periodic Timetabling: A SAT approach," 16th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems, 2016.

[16] Gattermann, P., Großmann, P., Nachtigall, K., and Schöbel, A. "Integrating Passengers' Routes in Periodic Timetabling: A SAT Approach," 16th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2016), vol. 54, pp. 3:1–3:15, 2016.

[17] Goverde, R. M. P. and Sparing, D. "A cycle time optimization model for generating stable periodic railway timetables," Transportation research, Part B. Methodological, 2017.

[18] Jovanovi, D. and Harker, P. T. "Tactical Scheduling of Rail Operations: The SCAN I System," Transportation Science, vol. 25, no. 1, pp. 46-64, 1991.

[19] Jovanovic, D. and Harker, P. T. "Tactical Scheduling of Rail Operations: The SCAN I System," Transportation Science, vol. 25, no. 1, pp. 46–64, 1991.

[20] Kümmling, M., Großmann, P., Nachtigall, K., Opitz, J., and Weiss, R. "A state-of-the-art realization of cyclic railway timetable computation," Public Transport, vol. 7, no. 3, pp. 281-293, 2015.

[21] Lamorgese, L., Mannino, C., and Natvig, E. "An exact micro–macro approach to cyclic and non-cyclic train timetabling," Pergamon, 2017.

[22] Li, X. and Meng, L. "Railway Scheduling Model under Temporary Speed Limits Based on Track Section Locking Time," Transportation Systems Engineering and Information, 2019.

[23] Liebchen, C. "The First Optimized Railway Timetable in Practice," Transportation Science, vol. 42, no. 4, pp. 420-435, 2008.

[24] Liebchen, C. "The First Optimized Railway Timetable in Practice," Transportation Science, vol. 42, no. 4, pp. 420-435, 2008.

[25] Liebchen, C. "Symmetry for Periodic Railway Timetables," Electronic Notes in Theoretical Computer Science, vol. 92, pp. 34-51, 2004.

[26] Lockwood, C. D. "A Model, Algorithms and Strategy for Train Pathing," Journal of the Operational Research Society, vol. 46, no. 8, pp. 988-1005, 1995.

[27] Luan, X., Corman, F., and Meng, L. "Non-discriminatory train dispatching in a rail transport market with multiple competing and collaborative train operating companies," Transportation Research Part C Emerging Technologies, vol. 80, no. jul., pp. 148-174, 2017.

[28] Kinder, M. "Models for Periodic Timetabling," master's thesis technische universität, 2008.

[29] Meng, L. and Zhou, X. "Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables," Transportation Research Part B Methodological, 2014.

[30] Liebchen, C. and Möhring, R. H. "The Modeling Power of the Periodic Event Scheduling Problem: Railway Timetables — and Beyond," Springer-Verlag, 2007.

[31] Nachtigall, K. and Opitz, J. "Solving Periodic Timetable Optimisation Problems by Modulo Simplex Calculations," 8th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS'08), vol. 9, pp. 1–15, 2008.

[32] Nie, L. and Zhang, Y. "Construction of Peak-Hour Periodic Train Timetables Using Breadth-First and Depth-First Search Methods," Railway Transport and Economy, 2013.

[33] Odijk, M. A. "A constraint generation algorithm for the construction of periodic railway timetables," Transportation Research Part B Methodological, vol. 30, no. 6, pp. 455-464, 1996.

[34] Petering, M. E. H., Heydar, M., and Bergmann, D. R. "Mixed-Integer Programming for Railway Capacity Analysis and Cyclic, Combined Train Timetabling and Platforming," Transportation Science, vol. 50, no. 3, pp. 892-909, 2016.

[35] Chen, J., Bierlaire, M., Robenek, T., Maknoon, Y., and Azadeh, S. "Passenger centric train timetabling problem," Transportation Research Part B Methodological, 2016.

[36] Robenek, T., Azadeh, S. S., Maknoon, Y., and Bierlaire, M. "Hybrid cyclicity: Combining the benefits of cyclic and non-cyclic timetables," Transportation Research Part C: Emerging Technologies, vol. 75, no. FEB., pp. 228-253, 2017.

[37] Rodriguez, J. "A constraint programming model for real-time train scheduling at junctions," Transportation Research Part B Methodological, vol. 41, no. 2, pp. 231-245, 2007.

[38] Serafini, P. and Ukovich, W. "A Mathematical Model for Periodic Scheduling Problems," SIAM Journal on Discrete Mathematics, vol. 2, no. 4, pp. 550-581, 1989.

[39] Wang, B. and Yang, H. "Periodic Potential Difference Model Based on CPF Constraint Graph," Journal of the China Railway Society, 2007.

[40] Xie, M. and Nie, L. "Sequencing-Based Periodic Timetable Model for High-Speed Railways Based on PESP," Transportation Engineering, 2009.

[41] Yan, F., Besinovic, N., and Goverde, R. M. P. "Multi-objective periodic railway timetabling on dense heterogeneous railway corridors," Transportation Research Part B: Methodological, vol. 125, no. JUL., pp. 52-75, 2019.

[42] Zhang, X. and Nie, L. "Integrating capacity analysis with high-speed railway timetabling: A minimum cycle time calculation model with flexible overtaking constraints and intelligent enumeration," Transportation Research Part C Emerging Technologies, vol. 68, no. jul., pp. 509-531, 2016.

[43] Bussieck, M. R., Winter, T., and Zimmermann, U. T. "Discrete Optimization in Public Rail Transport," Mathematical Programming, vol. 79, no. 1, pp. 415-444, 1997.

[44] Zimmermann, U. and Lindner, N. "New Perspectives on PESP: T-Partitions and Separators," OpenAccess Series in Informatics (OASIcs), vol. 75, pp. 2:1–2:18, 2019.

Research on the Application of Artificial Intelligence Product Design in Human Emotions: A Case Study of Chinese Women

Xiaolu Guo
College of Engineering, Design and Physical Sciences, Brunel university of London, London, The United Kingdom

Abstract—Human emotional and health issues have been the subject of world concern. In psychology, the correlation between them is indivisible, which is also a major challenge in studying the improvement of human emotions.
To address this challenge, this study aims to explore the impact of design on human emotional health, with China as the scope of the survey, and design an application to design prototypes and AI intelligent products. Quantitative and qualitative research methods are used, using questionnaire survey methods of quantitative research and literature quality research, to summarize and observe the data, to find user pain points; competing product analysis methods, to analyze existing market applications; UX design method based on Garrett's five-faceted method, using emotional analysis design prototypes, to idealize product content. The research direction of this paper is to immersive involvement through external intervention in human emotional behaviour through emotional design, improve the human psyche, thereby lining the stability of self-emotion, and enhance the subjective sense of human well-being.

Index Terms—Emotional well-being (EWB), Emotional design, Emotion computing , Artificial Intelligence, Product design

I. INTRODUCTION

As people age, People's concerns over their emotional health and well-being are growing. Weare, (2000) mentions that Emotional literacy refers to the capacity to recognize, comprehend, and apply knowledge about one's own and other people's emotional states. It can be seen that people can use cognitive processes and behavioural strategies to focus their emotional well-being.
However, the female group has challenges in maintaining emotional stability compared to the male group, mostly owing to variables such as job pressure, study pressure, and life pressure. Barrett and Toothman, (2016) indicated the value impact of ageist and sexist society on women's emotional well-being throughout their adult lives. Moreover, Women's subjective experiences, influenced by their self-perception, contribute to decreased enjoyment and increased health issues as a result of their concern around their selves. stated differently, negative attitudes may have an enduring impact on their mental well-being. In psychology, emotion regulation (ER) includes selection of the situation, modification of the situation, deployment of attention, change of cognitions and modulation of responses (Gross, 1998). Therefore, Individuals' cognitive-behavioral emotion control abilities are a means by which adults may preserve emotional stability.
User-Centered Design (UCD) can fulfill consumers' requirements and address their issues via external involvement. However, the primary emphasis of this research is on the individual's emotional well-being(Lowdermilk, 2013). According to Norman (2004), emotional design produces positive experiences and focuses on the user's emotions. Anderson (2011) also indicates that an application's visual design, which is intended to create good experiences and evoke emotions, has a direct impact on users' trust and sense of identity. Emotional design involves crafting the interface in a manner that captivates users via visually stimulating effects, hence leaving a lasting first impression. To address the abstract nature of human emotions, a prototype of a mobile application named U-MOOD was developed.

II. RELATED WORKS

The concepts of the ABC model of emotions, flow experience and the application of emotion computing can explain that there is a certain relationship between emotion quantification and big data of artificial intelligence, and explore the application of emotion recognition in product design. Affective computing is capable of detecting and examining human emotions. By assessing its conceptual depth of analysis, it explores the two components of affective identification and affective analysis. These components can serve as technological support for research purposes.

A. The ABC theory of emotion（ABC model）

Activating event, consequences and beliefs are three important components of the emotional ABC theory (see to figure 1). According to Albert Ellis's emotional ABC theory, when individuals experience negative emotions and fail to intervene promptly, it leads to psychological alterations that adversely affect their physical and mental well-being. The ABC model is an integral component of Rational Emotive Behaviour Therapy (REBT), employed to comprehend the mechanism of emotional distress. REBT technology, based on the ABC hypothesis, has been extensively utilised in non-treatment settings and implemented across several fields (Wu et al., 2021). Katsikis et al. (2016) clarify that Emotional ABC models may assess human emotions and cognitions and incorporate them into the design field. This approach not only

Xiaolu Guo is with the College of Engineering, Design and Physical Sciences, Brunel university of London, London, The United Kingdom (e-mail: xiaolu798123@gmail.com).

emphasises the emotional needs of the user but also ensures alignment between human cognition and behaviour. According to Denecke et al (2020), in the development of chatbot applications, the objective of emotional adjustment is accomplished by recognising emotional information and implementing applications such as robots, apps, and CBT mindfulness therapies. Therefore, the emotional ABC theory can assist consumers in recreating a belief environment or product that stimulates and fosters positive emotional needs and outcomes.

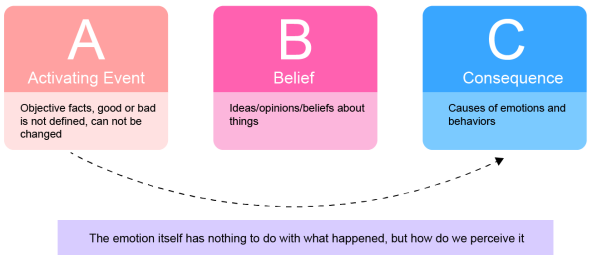


Figure 1: The ABC theory of Emotion

B. Flow experience

The flow experience is a progression that moves from being dependent on certain conditions, to being based on personal experience, and finally leading to a desired outcome (Nakamura and Csik Szentmihanyi, 2002). This illustrated the transformation of the flow phenomenon chart, expanding it from three experience regions to eight experience channels, these channels represent the varying levels of intensity experienced, depicted in a concentric ring. DeMatos, Sá, and Duarte (2021) mention that there is a correlation between subjective consciousness and emotions such as joy, pleasure, and loss of control when individuals engage in tasks or events. Emotions serve a significant role in emotional design by influencing an individual's decision-making, attention, and memory, thereby impacting the user's overall experience. Csikszentmihalyi's proposal encompasses nine conceptualised dimensions, (deMatos, Sa, and Duarte, 2021). Flow experiences provide a new focused user experience in emotional design, creating products and environments that balance challenge and skill, increased focus, clear goals, immediate feedback and an underlying sense of control under the key elements of flow experiences, thereby providing a sense of accomplishment and motivation to enhance their emotional well-being. Therefore, considering the double-sided nature of emotional design, the design will affect the user emotionally, and the user's emotions will have a certain view of the design. The application of flow experience makes the design not only meet the functional needs, but also contribute to the user's sense of well-being. Despite this, three conditions are required for flow to occur, namely that the challenges and skills should match the individual's abilities, that the user provide immediate, accurate feedback on the progress of the experience, and that the goals of the experience are clearly set (deMatos, Sa, and Duarte, 2021). Therefore, the understanding of the concept of flow experience and the application of flow experience in emotion design is very important for the design research of emotion management.

C. The utilisation of Affective computing

"Emotion computing", also known as Emotion AI, includes human emotions, emotions and feelings, emotion recognition and emotion analysis. It is also a branch of artificial intelligence that enables computers to interpret, simulate and react to human emotions (Wang et al., 2022). Emotion identification and emotion analysis are distinct aspects within the field of emotion computing. Emotion identification primarily involves identifying individuals' emotional state, with a focus on visual, linguistic, and physiological cues. On the other hand, emotion analysis is using technology to analyse social interactions, evaluate language, and determine good, negative, or neutral emotional outcomes (Balazs and Velasquez, 2016). This show that emotional computing has a wide range of applications in different fields, automatic recognition of emotions has changed the emotional information provided by human emotions, and the machine has expanded various fields to a greater extent. According to Madhusudan and A.K. (2016), emotion recognition refers to the act of obtaining, examining, and interpreting certain emotions displayed by users. Language, facial expressions, and emotion-related body postures can serve as means of conveying an individual's emotional state. Leong et al., (2023) mention that facial expression recognition for visual emotion recognition is achieved through three stages of machine registration, representation and recognition using pattern recognition, and facial expression data is easy to collect.

In this study, combining the emotion computing technology and machine algorithm, facial image analysis is carried out through the camera-based products to complete the analysis of facial expression recognition. Cai et al., (2018) state that face image analysis is performed by the camera, which necessitates specific conditions for shooting and where the data acquired varies with the intensity of light. This demonstrates that the design process takes into consideration the limitations and disadvantages of camera-based recognition of face emotions. Hence, this study primarily focuses on the utilisation of facial emotion recognition and expression in the realm of digital design.

III. Research Methodology

The research design used for this study is a qualitative approach that combines explanatory research with descriptive research aspects to elucidate the causal link between variables. Thus, an interpretive research methodology was employed to explore the manner in which design components convey human emotions and to contemplate strategies for mitigating adverse human emotions. And in the quantitative research design, the study of human emotions quantification is the difficulty of this study, designing the content of the survey, collecting a certain amount of samples, and analysing the quantitative content of emotions. Therefore, the purpose of this study is to understand the human emotions comprehensively in the process of designing and rationalising the improvement of emotional problems.

The literature search was initiated by doing a sample search using various combinations of keywords, resulting in the

identification of 77 items during the initial selection process. The content of the articles was cross-checked using the approach of literary Mata analysis. To mitigate the presence of redundant content from other databases, a thorough examination of the articles was conducted, resulting in the re-identification of 38 articles. Furthermore, the remaining articles yield 22 items through the assessment of their suitability based on the abstract and keywords. Upon carefully reviewing the article, limit the selection of relevant articles to a total of 14. Hence, a total of 14 items spanning from 2015 to 2023 have been chosen to be preserved for future study purposes. As depicted in table 1.
However, The sample for the questionnaire was obtained by releasing the questionnaire online for one month, resulting in a total of 82 valid responses. By analysing the initial data gathered on the web platform, it is possible to do an intuitive preliminary screening and conveniently determine if the quantity of questions fulfils the required criteria. Furthermore, the SPSS analytical technique was employed to establish the variables of the sample and perform a secondary analysis of the data in order to effectively and precisely achieve the research objectives.

IV. Data Analysis

Data was gathered from a total of 14 items carefully evaluated research papers and 82 questionnaires.

A. Findings from Qualitative Analysis

The study review of this article focuses on the significance of design content in relation to individuals' emotions, specifically examining the elements that influence emotions and the methods for regulating them. Furthermore, a cumulative of 18 scholarly articles were utilised for research purposes spanning from 2015 to 2023 (see to figure 2).

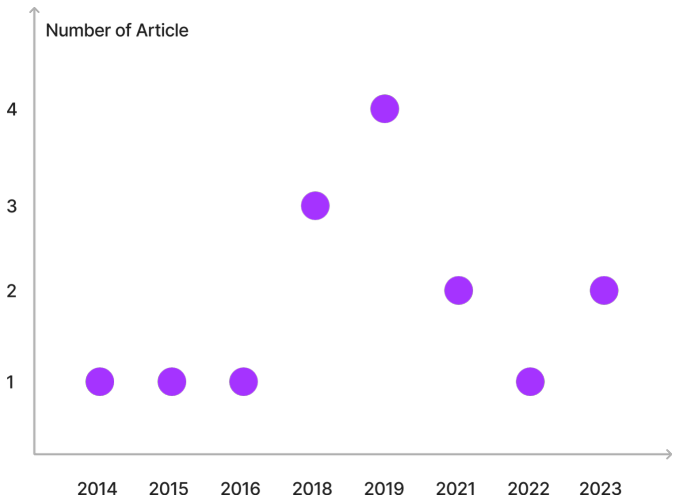


Figure 2: Number of articles published by year

1.Design elements affect people's emotional psychology

The development of science and technology has had an impact on people's emotional well-being as a result of the stress caused by multiple variables. Hence, by conducting a thorough examination of the article, several design components have the ability to evoke emotional reactions from users. The functional design of products and interfaces has the ability to alter consumers' emotions, thereby impacting their overall user experience. Moreover, as depicted in figure 3, it is evident that certain design components have a discernible impact on individuals' emotional psychology. Therefore, this study specifically examines the aspects of colour, shape, line, and function. This study specifically examines the aspects of colour, shape, line, and function.

Design element	Description of the influence of design elements and emotions
Color	Warm colours evoke sensations of warmth, comfort, enthusiasm, and other emotions, whereas chilly colours symbolise tranquilly, sadness, and apathy
Lighting	Intense illumination evokes a sense of positivity, whereas dim illumination induces a state of relaxation and can perhaps contribute to a melancholic ambiance
Space and Layout	The spatial arrangement of the design can influence individuals' perception of cognitive freedom or constraint
Shapes and Lines	Curvilinear forms and contours are frequently regarded as calming and organic, whereas sharp, angular lines can elicit a perception of structure and exactitude, but may also convey a sense of aggression
Texture	Texture has the ability to communicate feelings of both comfort and discomfort. It can enhance the visual experience by adding a tactile element and has the power to affect emotional reactions
Symmetry and Balance	Symmetrical designs are commonly seen as harmonious and aesthetically pleasant, whereas asymmetrical designs can be more dynamic and enjoyable, but may occasionally evoke emotions of disquiet or imbalance
Typography	The choice of font style in the design might also elicit an emotional response
Patterns and Repetition	Patterns can engender a feeling of stability and predictability, so providing solace. Excessive repetition, on the other hand, might become tedious and uninteresting
Personal and Cultural Associations	Human experience and cultural background can significantly impact emotional reactions to many design aspects
Context and Functionality	The mood of a design can be influenced by both its setting and its function

Figure 3: The impact of design aspects on emotional psychology

2.Methods to regulate emotions

Emotional regulation is a key characteristic of mental well-being. Hence, it is crucial to have a diverse range of ways to effectively regulate unpleasant emotions, such as music, engaging in physical activity, practicing deep breathing, seeking social support, smiling and so on. In Facial Expression Training, since there are variations in emotional reactions and perceptual acuity, individuals may be less perceptive to discerning face expressions. Individuals are unable to perceive their own facial expressions, and the primary method to enhance face expression is through practicing expression. When self-training facial expressions, some external media intervention is needed. Han et al., (2023) pointed out that the poorly supervised DECD framework utilises static facial expression photos for training in order to detect the timing of emotional changes, due to certain issues in computer vision analysis.nTherefore, for an effective facial expression training system, it is necessary to provide user selectable target facial expression interface. Training to change facial expressions from negative to positive or neutral can be interfered with using a number of methods (see to figure 4). Among them, mirror exercise, expression imitation and facial feedback hypothesis are effectively used in this study.

Method	Paraphrase
Awareness Training	Being conscious of one's own emotions and facial expressions is typically the first step. Practices like mindfulness or meditation that help people to notice their emotions and expressions without passing judgement can help achieve this.
Mirror Exercises	People can focus on changing their facial motions and become more aware of them by practicing in front of a mirror. For instance, when remembering a tense situation, one could make an effort to keep a neutral demeanour.
Biofeedback	Biofeedback devices can assist individuals in identifying negative emotions and gaining control over their reactions. Your concept's intelligent recognition mirror is a kind of biofeedback equipment that offers real-time analysis and adjustment suggestions.
Relaxation Techniques	Deep breathing, progressive muscular relaxation, and visualisation are among techniques that can be used to manage emotional arousal, which frequently comes before negative reactions.
Expression Imitation	Seeing pictures or videos of various face expressions and trying to mimic them can aid in improving facial muscle control.
Emotion Regulation Strategies	The underlying emotions that give rise to unpleasant expressions can be altered by the application of cognitive-behavioral techniques. This may be rephrasing unfavourable ideas or concentrating on the good things that have happened.
Acting Techniques	Performers frequently train their facial expressions to convey a range of emotions. Individuals can learn to control their own expressions by adapting some of these approaches.
Professional Therapy	Working with a therapist can occasionally assist individuals in identifying the underlying causes of their negative expressions and creating coping mechanisms.
Social and Emotional Learning (SEL)	Through this process, people can learn how to identify and control their emotions, create and meet constructive goals, and empathise with others.
Facial Feedback Hypothesis	This idea says that facial gestures might be able to change how people feel. One example is that smiling on purpose can make someone feel better.

Figure 4: Methods of facial expression training

B. Findings from Quantitative Analysis

The questions in the questionnaire are of various compound types, such as multiple choice questions, multiple choice questions and open-ended questions, so as to avoid the multiplicity of content. The participants selected for the survey are mainly Chinese adults aged 18 to 65. According to the survey, people in this age group can operate the basic functions of smart phones, so the survey participants can easily access and fill in the questionnaire. A total of 86 valid samples (n = 86) were obtained for the initial evaluation of the questionnaire. A cross-analysis was conducted by selecting several dependent and independent variables. The data regarding the participants' gender, age, and education were condensed in order to establish the fundamental characteristics of the participants.

1. Reliability analysis

Reliability analysis is a research method to study whether the quality of data is reliable. A total of 86 subjects were included in the study, with a total of 67 voluntary participants covering 20 programs. Through Cronbach. α alpha analysis of the data, as shown in the table 1, the reliability coefficient of the scale was obtained as 0.996, indicating that the scale had high reliability and internal consistency. This outcome further confirms the dependability and consistency of the scale in assessing target variables, and offers trustworthy foundational data for ongoing study.

Sample size (N)	Number of items	Cronbach's Alpha
86	20	0.996

Table 1 : Reliability analysis

2.Validity analysis

Validity is a metric employed to assess the rationality of the design of a given item, specifically in the context of quantitative data and limited to scale data. To accomplish the desired analysis, a total of 12 questions were chosen from a pool of 14 questions. One of the questions pertained to quantifying emotional words, while the remaining 10 questions were expanded. After in-depth analysis of the data generated, the questionnaire validity analysis mainly involves two parts: factor analysis and Kaiser-Meyer-Olkin (KMO) test. In the factor analysis part, 10 factors were selected from the questionnaire, and the variance explanation rates of these factors ranged from 1.00% to 67.56%. The cumulative variance explanation rate before rotation is 96.92%, and the cumulative variance explanation rate after rotation is 96.92%. The eigenroot values before and after rotation are all greater than 1, indicating that each factor can explain part of the variability in the questionnaire. The variance interpretation rate after rotation is lower than that before rotation, but still presents a higher interpretation rate overall. The result of KMO test was 0.945, close to 1, indicating that the questionnaire was very suitable for factor analysis. The value of Barth's sphericity test was 4704.665, and the significance P-value was less than 0.05, which further proved that the questionnaire was suitable for factor analysis. Therefore, the questionnaire has high validity and good structural validity, which is suitable for follow-up research and analysis in emotional health, psychology and other aspects.

3.Comparative analysis

Comparative analysis can be used to set one or more independent variables and dependent variables, so as to obtain the difference of dependent variable data at different levels of independent variables, and presented in data tables or line charts, bar charts, etc. Therefore, the correlation between the questionnaire questions was analyzed for the set independent variables and dependent variables. In addition, for the convenience of viewing, the serial numbers of the questions were marked T1, T2, T3... to T14. When the independent variables are T1 (X= gender) and T2 (X= age), and the dependent variables are T5, T6, T7, T8, T10, T11, and X=T1, T2, Y=T10 in-depth analysis is performed.

Hypothesis 1: X=T1, T2, Y=T5.

Based on the cross-analysis of anxiety frequency among different age groups and genders, as shown in Figure 5, it can be observed that women between the ages of 18 and 35 experience a higher level of anxiety. On average, these women experience anxiety episodes every three to seven days. The findings indicate that women experience a notable level of anxiety, and the frequency of the anxiety cycle is higher.

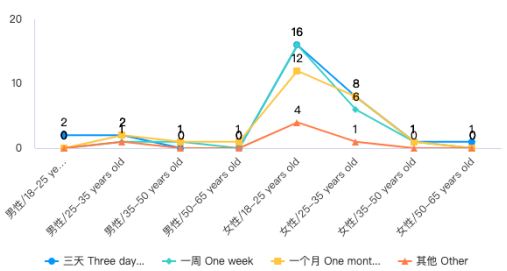


Figure 5: Anxiety's recurring pattern

Hypothesis 2: X=T1, T2, Y=T7.

According to Figure 6, the participants were distributed as female groups, while the longitudinal analysis.The way to relieve anxiety, playing mobile phones, listening to music is the preferred way, followed by reading, sports, travel. From this point of view, people have a high dependence on electronic

products, and the way to choose mitigation is not the traditional form, which also shows that people choose products under digital intelligence to intervene, because time or economy is not allowed, so they choose ways that can be used frequently in daily life to alleviate.

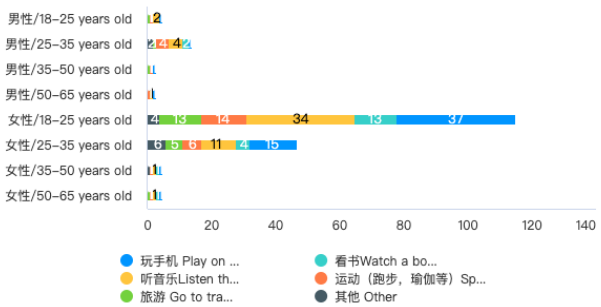


Figure6: A way to ease your emotions

Hypothesis 3: X=T1, T2, Y=T8.
Based on the longitudinal analysis chart (see figure 7), more than a third of the participants were never physically or psychologically examined. According to the data, the sample of male participants is small, but basically tend to pay attention to physical health or mental health, while the female participants in the age of 18-35 are mostly not concerned. This shows that women are affected by different environmental factors that affect their health, resulting in a lot of health or emotional problems, so it is difficult to find emotional health problems.

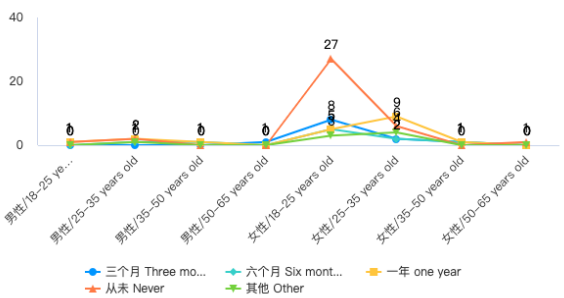


Figure7: The frequency at which participants assessed their own physical and mental well-being

Hypothesis 4: X=T1, T2, Y=T10
According to figure 8, out of the 86 participants, 76 have not utilised emotion management applications. This suggests that either consumers are not interested in a particular product or they lack awareness of similar products. Nevertheless, studies indicate that individuals experience enhancements when utilising emotion management applications, goods, or employing other methods to interfere in their emotions. The data collected in the questionnaire serves as guidance for the ensuing design process, aiming to create products or novel designs that captivate people' attention.

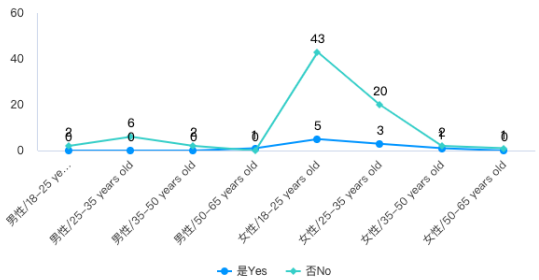


Figure8: Participants' use of an emotion management APP

Hypothesis 5: X=T1, T2, Y=T11.
As can be seen from Figure 9, participants' understanding of the correlation between emotions and health is relatively obvious, but the degree of understanding only tends to a general level. This also shows that people only pay attention to a topic of emotional problems, and have no initiative to deeply understand them.

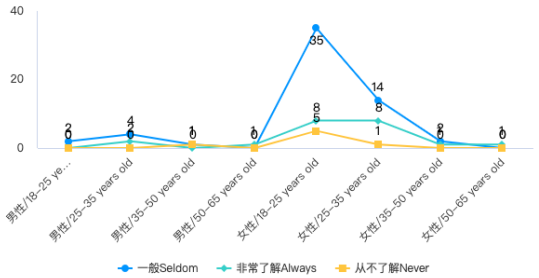


Figure 9: The participants' level of understanding of emotional health

Hypothesis 6: X=T7, Y=T11, T12
Figure 10 and figure 11 demonstrates the cross-analysis between voluntary participation in the emotional test and the educational level of participants when examining emotional health. According to the data, the cohort beyond the college level possesses a comprehension of emotional well-being and actively engages in the examination. These findings indicate a correlation between individuals' level of education and their concern for emotional well-being.

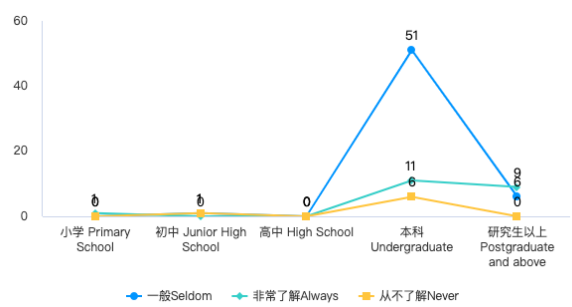


Figure10: To understand the relationship between emotional health and the educational level of the participants

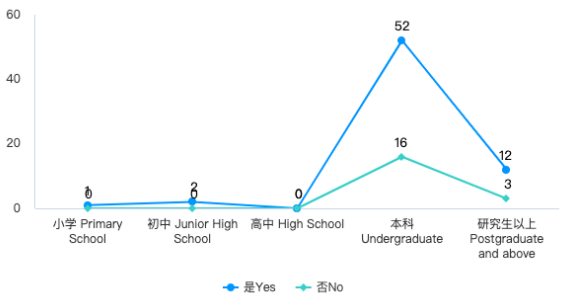


Figure 11: Relationship between participation in mood tests and participants' educational attainment

V. Design and Evaluation

Utilising APP (UMOOD) design and product design as the primary mediums. The main goal of this chapter is to apply design thinking, specifically focusing on the four stages of Design thinking (Empathize/Define, Ideate/Prototype, Design/Build, and Review/Refine).

A. Target user

It has been determined that individuals between the ages of 18 and 35 have a keen interest in the aforementioned issues, and their emotions are influenced by numerous factors. Hence, the intended demographic for this study comprises individuals aged 18 to 35(see to figure 12).

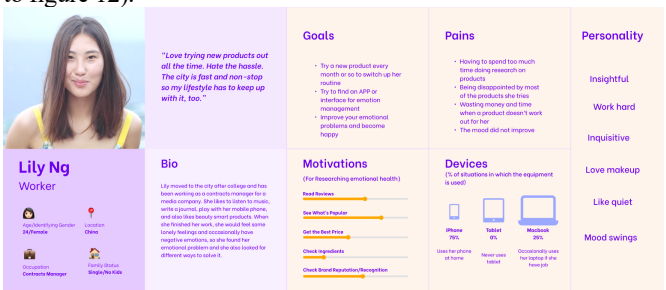


Figure 12： Persona

B. Problem Definition

Around the 5W1H (i.e, what, why, where, who, when, how) principle is defined, from external factors to internal factors, in-depth analysis of the problems faced by users, to find users' pain points and needs. The main problem defined in this study (see to figure 13) is that users lack the vision to detect emotions and lack some sense of self-affirmation in their hearts. As for the main problem, the solution is that it is difficult to change due to external factors. The user's inner psychological self-awareness builds the awareness of discovering emotions, and influences people's emotions through design to achieve a stable positive emotional state.

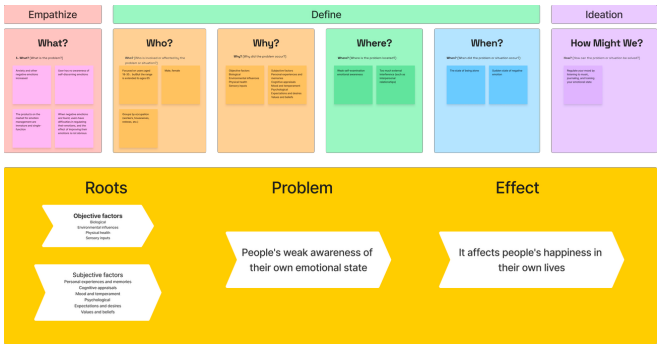


Figure13： Define problem

C. Emotion vision design

Emotion is an abstract concept that is visually represented via the use of colour and line in the design elements. The emotion data is then incorporated based on the quantification of emotions obtained from the questionnaire. Figure 14 illustrates the usage of warm colours to represent happy emotions, whereas cool colours are employed to depict negative emotions. The closed shape of the emotion visualisation pattern is displayed. The central motif is encircled by a circular dial, indicating the specific moments when emotions are manifested, and allowing for a clear observation of emotional fluctuations. This design prepares for the subsequent product design. Under the current technological feasibility, the quantified emotion data table is visualized and input into the product. Through the big data of AI, it is analyzed based on factors such as users' personal usage and environmental changes, thereby obtaining a more rigorous data presentation.

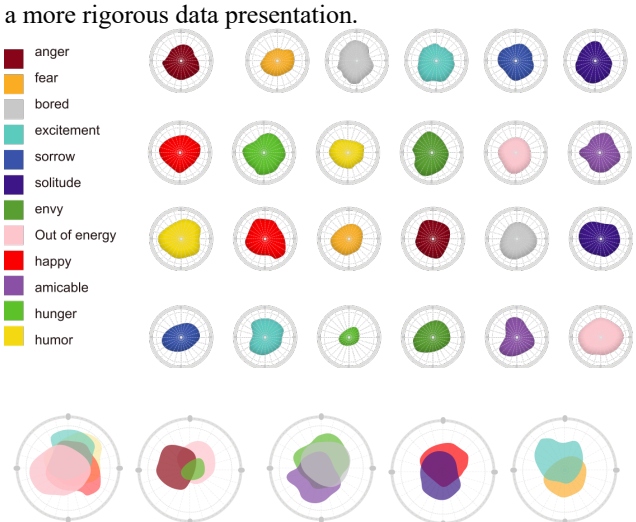


Figure 14: Emotion vision design

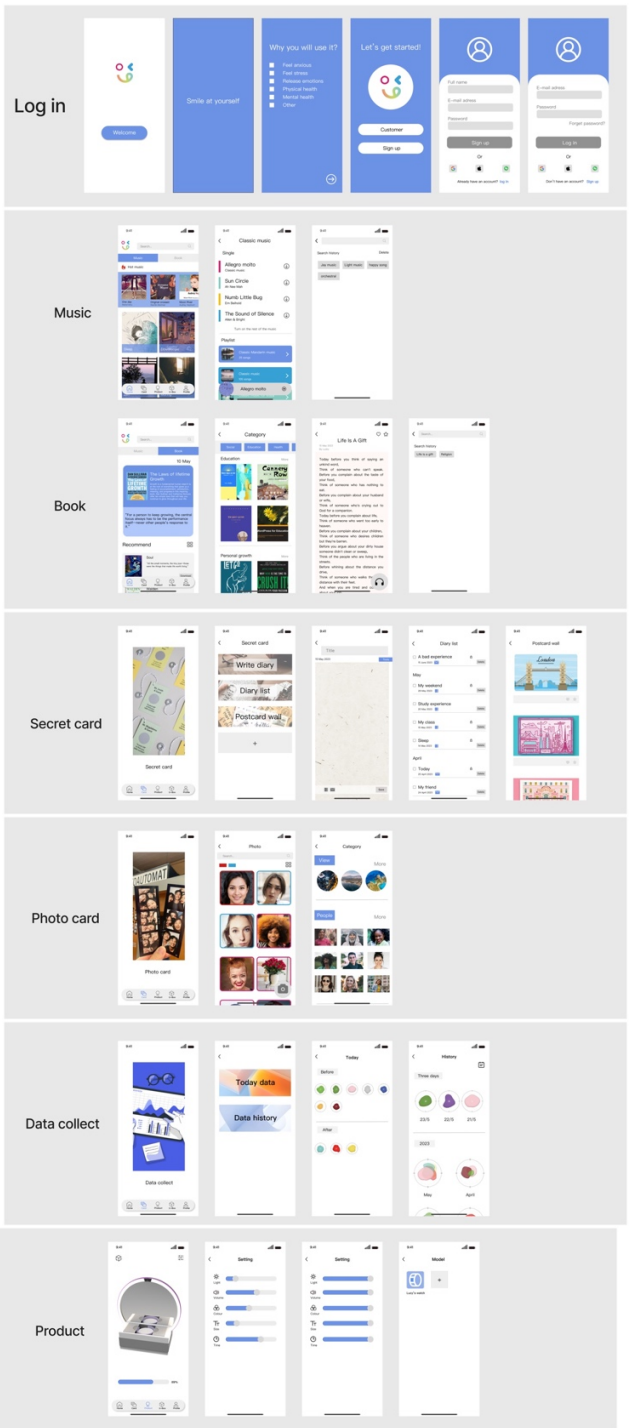
D.APP design

The main function of the application (APP) is to collect product data, analyze it through big data and technology, and collect users' emotional changes for visual display. This function is required to be used in conjunction with the product. The detailed introduction of the APP is as follows:

1. Following the literature research and data analysis, the App utilises the mood therapy approach by suggesting music and proposing articles or books. The selection of these materials

takes into account their suitability for promoting happy feelings. The primary criterion for music selection is predominantly light music, such as the cello. Furthermore, music of comparable genres is further categorised, and the colour of each music entry also varies based on the distinct emotional states evoked by each piece. The article selection criteria are characterised by their positivity and readability, exemplified by works such as "The Little Prince."

2. The auxiliary products, including the bracelet module, can be shown and their fundamental functions can be altered through mobile phone operation. This includes modifying the light source of the product and setting the duration of usage. Furthermore, the functionality of the product can be assessed to ascertain its usability and effectiveness in serving as a reminder for users.
3. The function of the card is to gather the user's personal content, which can enhance their bad emotions and provide insights about mood fluctuations. This can be achieved by customising the card content or utilising pre-set cards. The literature research and analysis have summarised one of the three strategies of emotional regulation, which is creative activity through writing. Fixed cards typically consist of three primary elements, one of which is the privacy card. This card enables users to document their daily mood by writing down fragmented text on a regular basis. Additionally, the photo card primarily gathers up-to-date facial photographs of users throughout product usage. Furthermore, the task involves categorising and exhibiting the photographs, essentially creating a photo repository. The data collecting and analysis card primarily displays the emotional changes that have been analysed using emotional computing technology in products and apps. Furthermore, the emotional visualisation feature showcases the user's emotions within a given time frame, which can be as precise as an hour interval. The data history includes a 24-hour record, a three-day record, and a monthly record.
4. The function of the U-box is to facilitate effective communication between users and various groups of individuals. Research has indicated that when individuals experience negative emotions in response to external stimuli, the majority of people tend to withdraw and limit their interactions with others. This is due to individuals exhibiting adverse behaviours towards social engagement and self-expression, which hinders emotional well-being. Furthermore, the online contact facilitated by the Internet undermines the robustness of face-to-face communication, hence providing significant assistance to individuals who experience communication deficiencies. The objective of this study is to ensure that consumers consistently maintain a steady emotional state and bolster their confidence. Hence, inside the design framework of U-box, individuals have the option to engage in communication with various groups, encompassing close connections such as family and friends, as well as unfamiliar individuals. When establishing this communication function, the user information is supposed to be secret, and the particular details are not shared between users. However, a broad description is provided. Furthermore, certain demographic segments exhibit heightened emotional distress, necessitating an augmented reliance on private healthcare practitioners. The combined impact of the application (APP) and the product allows for the initial extraction of huge data for preliminary analysis. Subsequently, communication with a private doctor enables the doctor to gain insight into the user's daily emotional state, facilitating the development of an effective treatment plan. This function is discretionary and potential research.
5. Personal Settings refer to the customisation options available for modifying both personal information and the application itself.



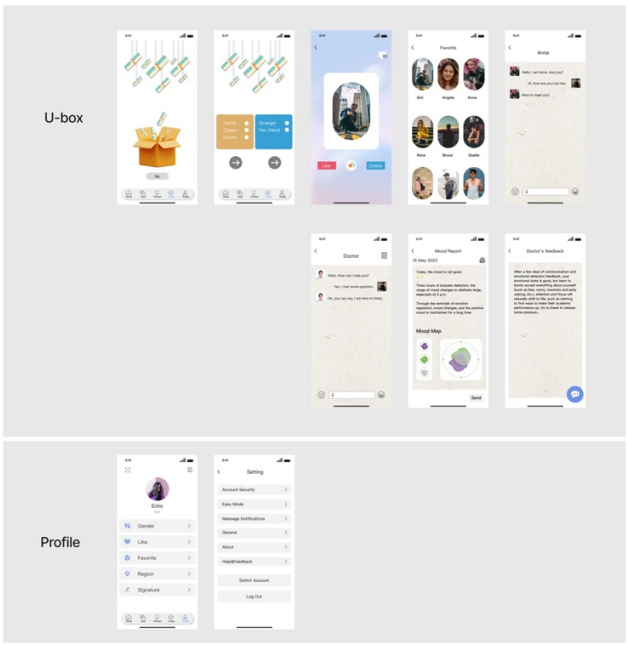


Figure 15:APP Design

E. Product design

According to market research and analysis, the integration of APP and product design is a strategy that enhances user experience and is frequently taken into account by designers. The product design in this study consists of a mirror and a smart bracelet that possess the ability to accurately detect and analyse emotional fluctuations. The product's functional design primarily consists of five functions. The analysis provided is comprehensive and thorough:

1.Through literature research and analysis, it is possible to apply emotional computing technology to products in order to detect and recognise users' emotions. By conducting a thorough analysis, one can receive data about feedback and recognition. If the user experiences prolonged or severe negative emotions, the product can serve as a reminder for the user to engage in self-training.

2. The product is used by the associated programme (UMOOD), which has the capability to play music and listen to the content of articles, among other things. The product design incorporates acoustic functionality. The objective of this design is to provide users with a means of escaping from the intricate information on their mobile phones during moments of negative emotions, and to allow them to indulge in the gratification and assurance offered by the immersive product.

3. Because the product is set as a smart makeup mirror, it can not only improve people's mood, but also act as an ordinary mirror to use, users can make up and so on. In addition, the design with a light belt can not only brighten the surrounding dark environment, but also be clearly identified in the smart mirror in recognizing people's emotions.

4. The bracelet is equipped with an intelligent recognition mirror as a component of the product. The lower compartment of the box is specifically intended to contain a concealed area that serves as both the charging and storage unit for the smart bracelet.

5. Intelligent mirror recognition primarily focuses on the impact of people's emotions. When intervention from external sources is necessary, users can receive feedback on changes in their facial emotions. The duration of emotional training takes precedence in identifying the user's emotions, and it is determined through internal technical analysis. Additionally, the length of emotional training can be customised.

6. The smart bracelet is mainly used to detect emotions. Because the use environment of the smart mirror is limited, it is not suitable for carrying. Therefore, the bracelet can always detect the user's emotional state, and the function of the bracelet is a single function, mainly for time viewing, movement distance and emotional function detection and feedback.

F. Product size

The size of the product is determined by considering the dimensions of similar genuine products and analysing the user's usage patterns and environment, in order to create a suitable and practical size for the product. Therefore, the products of this study are divided into a mirror that intelligently identifies emotions and a bracelet that intelligently detects emotions. The initial prototype has been established. The dimensions of the smart mirror's box component are as follows: 150mm in length, 100.03mm in width, and 60mm in height. The wristband storage space has dimensions of 70mm (length) x 56mm (width) x 32mm (height), and the mirror is a component inserted in the box with a radius of 94mm, the detailed dimensions are shown in Figure 16. The smart bracelet has a circular shape with a circumference of 300 degrees, a radius of 25.5mm, a width of 25mm, and a thickness of 1.5mm, the precise dimensions are depicted in Figure 17.

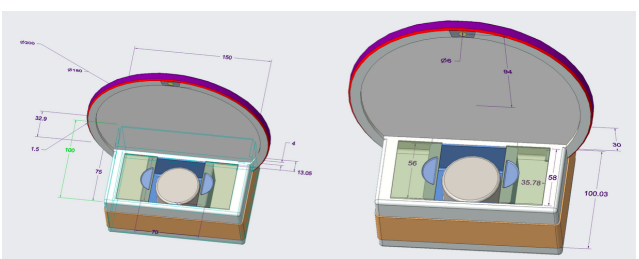


Figure 16: Smart mirror size

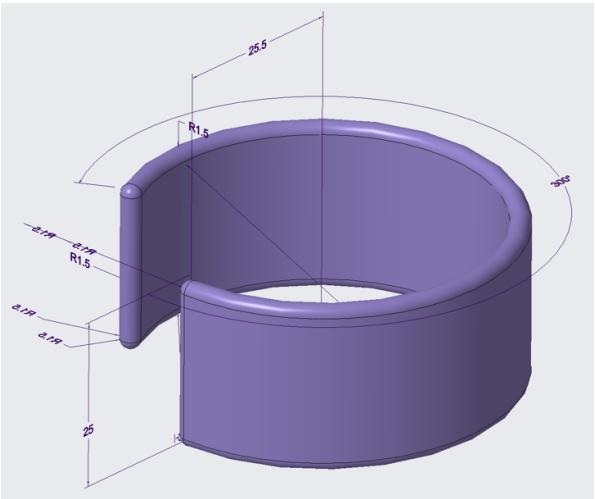


Figure 17: Smart bracelet size

The Hi-fi of the product is the display of the material of the product model, which can be close to the real product. In the design process of the product, the overall design of the product color selection is simple milky white. The sound material in the smart mirror product is fabric texture, and the switch of the storage bracelet is acrylic material (see to figure 18). The band is made of carbon fiber and the screen is made of glass (see to figure 19).

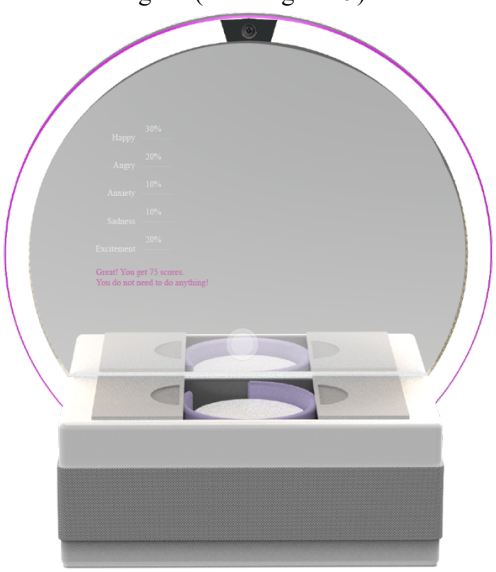


Figure 18: Smart mirror

Figure 19: Smart bracelet

Figure 20: Real scenes

G.Service blueprint

Before use, in use, and after use are the three phases that are included in a service blueprint (see to figure 21). This blueprint is a study of products and applications in conjunction with the three phases before and after use. Users have a clear understanding of the product and application service process since it is broken down into front-end user operations, back-end application analysis, and product analysis.

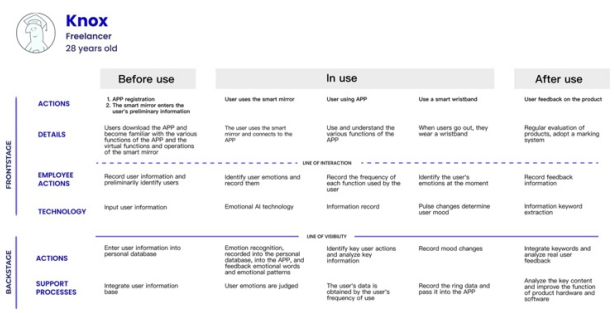


Figure 21: Service blueprint

H. Heuristic Evaluation

Heuristic evaluation is a method in which specialists assess the usability of user interfaces by applying established guidelines during individual walkthroughs and documenting any identified problems (Interaction Design Foundation,

2019). Based on ten usability heuristic methods to expand (see figure 22). Therefore, the evaluation of the design aspect of the study is conducted using the Nielsen-Molich method, which involves assessing the product based on a predetermined set of inspection criteria.

1. The APP integrates a smart mirror and bracelet to provide users with mutual reminders about the status of the device. For instance, mobile phone applications can effectively control the functionality of smart mirrors and wristbands.
2. Optimise the design of the APP based on a study of pertinent research on rival applications and consumers' usage patterns. As an illustration, the design incorporates a layout where three photographs are arranged horizontally to enhance user comprehension of the material. The product utilises a reduced language, employs graphical and textual representations for the APP icon, and avoids redundant repetition of fixed text information within the APP. The size and style of smart mirror goods are determined by measuring their dimensions and height, ensuring that people may utilise them more ergonomically and healthily.
3. The APP, smart mirror, and bracelet are primarily designed for interaction. Users can modify and personalise these devices according to their habits. For example, the content displayed on the APP cards can be customised.
4. Users can adapt their behaviour to preserve operational flexibility. Simultaneously, users can promptly get the needed information using the search function. During the initial usage of the smart mirror application, users can become acquainted with its functions through the automated design of the operation process. Additionally, users may encounter difficulty in navigating the settings function within the application.

Consequently, the three gadgets (APP, smart mirror, bracelet) have a mutual interaction, making the product design accessible.

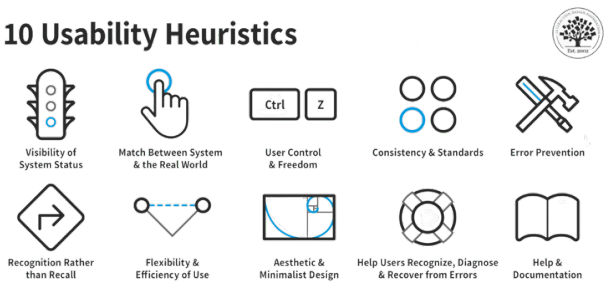


Figure 22: Ten usability heuristic methods
Source:(Interaction Design Foundation, 2019)

VI. Conclusion

This study examines the emotion visualization model, which is developed via data analysis. The objective of this study is to enhance individuals' emotional well-being by integrating an application (APP) with a product. In literature studies and surveys, people's awareness of emotional health issues is weak, while in market research, people's emotions are mostly analyzed and improved in the form of a single application. Thus, this study considers not only the principles of user experience but also the visually pleasant interface design of products and applications. By utilizing the developed products of this study, users can successfully enhance their emotions. The advantage of the design is that users can have an immersive environment to improve their emotions, and emotional feedback can be obtained through emotion recognition, so that users can self-select the three ways to regulate their emotions.

In psychology, emotions are intangible and challenging to quantify. The primary objective of this study is to measure emotions and propose optimal remedies using advanced technology. By searching keywords (emotion visualization, emotion visualization), the relevant literature summary is obtained. Through detailed reading and analysis, it is found that there is a lack of practical application demonstration in design in previous studies. By conducting a keyword search on emotion visualisation, the relevant literature summary is obtained. Thorough examination and analysis reveal a dearth of practical application demonstrations in design in prior research. Prior to the widespread adoption of products and apps, emotional assessment tables can be employed to gather users' current emotional states, establish a comprehensive database, and present abstract visual representations using design components such as colours and graphics.

The study not only innovatively captures facial expressions to communicate emotions, but also collects quantitative data on emotions. This research contributes to the field of combining psychology and design, particularly in the realm of digital design. This study investigates the research direction of facial expression training as a technique of adjusting emotions when receiving feedback. Once emotional AI technology reaches maturity and is implemented in practical settings, further study is required to accurately record individuals' emotional fluctuations and align them with their own feelings. Not only consider the theoretical support, but also need to consider the initial output of the design. This study requires long-term practical testing to obtain data and accurate detection rules. In the application design, the function to involve the psychiatrist in people's everyday lives, not just the traditional way of consultation, in the case of self-management, when the user needs help, through the product's preliminary data method to give the doctor's initial judgment, in addition, for the person's emotional visualization patterns, can also under 3D printing technology, transform the emotional model into physical.

There are also limitations in this study, when collecting questionnaire data, the survey carried out only on the population of China, although the data obtained is effective, can detect problems and support the topic of the study, but the amount of data is not large enough, which is not enough for the accuracy of the research. Furthermore, in this study, it is difficult and limited to apply design applications and products to real life, not only taking into account the technology in the product, but also the organization of big data. For these constraints, the design of this study is reasonable, based on previous research and analysis and experience of market products and applications. In future studies, it is given priority. Secondly, according to the survey, there is a lack of awareness about improving emotional health, and the market products do not attract users, so attracting the attention of users also needs to be considered.

REFERENCES

[1] Alharahsheh, H.H. and Pius, A. (2020) 'A review of key paradigms: Positivism VS interpretivism', Global Academic Journal of Humanities and Social Sciences, 2(3), pp. 39-43.

[2] Babich, N. (2022) Low fidelity vs. high fidelity: the differences between design prototypes | Webflow Blog. [online] Available at: https://webflow.com/blog/low-vs-high-fidelity. . Available at: (Accessed: .

[3] Balazs, J.A. and Velásquez, J.D. (2016) 'Opinion Mining and Information Fusion: A survey', Information Fusion, 27, pp. 95-110 Available at: 10.1016/j.inffus.2015.06.002.

[4] Cai, Y., Guo, Y., Jiang, H. and Huang, M. (2018) 'Machine-learning approaches for recognizing muscle activities involved in facial expressions captured by multi-channels surface electromyogram', Smart Health, 5-6, pp. 15-25 Available at: 10.1016/j.smhl.2017.11.002.

[5] Cyr, D. (2014) Emotion and website design. Interaction Design Foundation - IxDF. Available at: (Accessed: .

[6] de Santana, M.A., de Lima, C.L., Torcate, A.S., Fonseca, F.S. and dos Santos, W.P. (2021) 'Affective computing in the context of music therapy: a systematic review', Research, Society and Development, 10(15), pp. e392101522844.

[7] deMatos, N.M.d.S., Sá, E.S.d. and Duarte, P.A.d.O. (2021) 'A review and extension of the flow experience concept. Insights and directions for Tourism research', Tourism Management Perspectives, 38, pp. 100802 Available at: 10.1016/j.tmp.2021.100802.

[8] Denecke, K., Vaaheesan, S. and Arulnathan, A. (2020) 'A mental health chatbot for regulating emotions (SERMO)-concept and usability test', IEEE Transactions on Emerging Topics in Computing, 9(3), pp. 1170-1182.

[9] Desmet, P.M.A., Porcelijn, R. and van Dijk, M.B. (2007) 'Emotional Design; Application of a Research-Based Design Approach', Knowledge, Technology & Policy, 20(3) Available at: 10.1007/s12130-007-9018-4.

[10] Ding, M. and Dong, W. (2019) 'Product color emotional design considering color layout', Color Research & Application, 44(2), pp. 285-295.

[11] Evans, J.R. and Mathur, A. (2018) 'The value of online surveys: A look back and a look ahead', Internet research, 28(4), pp. 854-887.

[12] Fancourt, D., Garnett, C., Spiro, N., West, R. and Müllensiefen, D. (2019) 'How do artistic creative activities regulate our emotions? Validation of the Emotion Regulation Strategies for Artistic Creative Activities Scale (ERS-ACA)', PloS one, 14(2), pp. e0211362.

[13] Giasiranis, S. and Sofos, L. (2017) 'Flow Experience and Educational Effectiveness of Teaching Informatics using AR', Educational technology & society, 20(4), pp. 78-88.

[14] Gutounig, R., Goldgruber, E., Ausserhofer, J., Andrews, K., Traunmüller, T. and Wolkinger, T. (2016) 'The Styrian diversity visualisation project: Communicating data stories with an open data visualisation web app', Tagungsband des 10.Forschungsforum der Österreichischen Fachhochschulen, .

[15] Haertl, K.L. and Ero-Phillips, A.M. (2019) 'The healing properties of writing for persons with mental health conditions', Arts & Health, 11(1), pp. 15-25.

[16] Han, B., Yoo, C., Kim, H., Yoo, J. and Jang, J. (2023) 'Deep emotion change detection via facial expression analysis', Neurocomputing, 549, pp. 126439 Available at: 10.1016/j.neucom.2023.126439.

[17] Hayes, M. and Hefferon, K. (2015) '‘Not like rose-tinted glasses… like taking a pair of dirty glasses off’: A pilot intervention using positive emotions in expressive writing.', International Journal of Wellbeing, 5(4), pp. 78.

[18] He, R., He, X., Su, Y., Wang, Y., Liang, T., Cui, Z. and Zhang, L. (2023) 'Effect of ABC Theory Model on Negative Emotion of Young Patients with Breast Cancer During Treatment', Journal of multidisciplinary healthcare, 16, pp. 1883-1888 Available at: 10.2147/JMDH.S405564.

[19] Inc, G. (2022) 'Global emotions report. [online] Gallup.com. Available at: https://www.gallup.com/analytics/349280/gallup-global-emotions-report.aspx.', .

[20] Inglehart, R. (2010) 'Faith and freedom: Traditional and modern ways to happiness', International differences in well-being, 351, pp. 397.

[21] Interaction Design Foundation (2020). Putting Some Emotion into Your Design – Plutchik’s Wheel of Emotions. [online] The Interaction Design Foundation. Available at: https://www.interaction-design.org/literature/article/putting-some-emotion-into-your-design-plutchik-s-wheel-of-emotions.

[22] Interaction Design Foundation (2019). What is Heuristic Evaluation? [online] The Interaction Design Foundation. Available at: https://www.interaction-design.org/literature/topics/heuristic-evaluation.

[23] Ip, P.K. (2011) 'Concepts of Chinese folk happiness', Social Indicators Research, 104, pp. 459-474.

[24] Jmour, N., Masmoudi, S. and Abdelkrim, A. (2021) 'A new video based emotions analysis system (VEMOS): an efficient solution compared to iMotions Affectiva analysis software', Adv.Sci.Technol.Eng.Syst.J, 6, pp. 990-1001.

[25] Joshanloo, M. (2014) 'Eastern conceptualizations of happiness: Fundamental differences with western views', Journal of happiness studies, 15, pp. 475-493.

[26] Kahn, J.H., Ladd, K., Feltner-Williams, D.A., Martin, A.M. and White, B.L. (2022) 'Regulating sadness: Response-independent and response-dependent benefits of listening to music', Psychology of Music, 50(4), pp. 1348-1361.

[27] Kandel, B. (2020) 'Qualitative Versus Quantitative Research', Journal of Product Innovation Management, 32(5), pp. 658.

[28] Katsikis, D., Kostogiannis, C. and Dryden, W. (2016) 'Rational-emotive behavior approach in life coaching', Journal of evidence-based psychotherapies, 16(1), pp. 3-18.

[29] Krause, A.E., Pardon, M., Hoang, M. and Lucano, R. (2023) 'Listen Up: A case study examination of focused listening', Musicae Scientiae, , pp. 10298649231203628.

[30] Langeland, E. (2022) 'Emotional well-being'Encyclopedia of quality of life and well-being research Springer, pp. 1-3.

[31] Leong, S.C., Tang, Y.M., Lai, C.H. and Lee, C.K.M. (2023a) 'Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing', Computer Science Review, 48, pp. 100545 Available at: 10.1016/j.cosrev.2023.100545.

[32] Leong, S.C., Tang, Y.M., Lai, C.H. and Lee, C.K.M. (2023b) 'Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing', Computer Science Review, 48, pp. 100545 Available at: 10.1016/j.cosrev.2023.100545.

[33] Lin, J. and Li, J. (2016) An effective user centered approach: Using web design framework to support user experience design of interactive multi-functional product. pp. 3129.

[34] Madhusudan and A. K. Sharma (2016) Affective computing: Emotion sensing using 3D images. pp. 150.

[35] Mesurado, B., Vidal, E.M. and Mestre, A.L. (2018) 'Negative emotions and behaviour: The role of regulatory emotional self-efficacy', Journal of adolescence, 64, pp. 62-71 Available at: 10.1016/j.adolescence.2018.01.007.

[36] Micheli, P., Wilner, S.J., Bhatti, S.H., Mura, M. and Beverland, M.B. (2019) 'Doing design thinking: Conceptual review, synthesis, and research agenda', Journal of Product Innovation Management, 36(2), pp. 124-148.

[37] Nakamura, J. and Csikszentmihalyi, M. (2002) 'The concept of flow', Handbook of positive psychology, 89, pp. 105.

[38] O'Donoghue, J. (2022) Why Are Personas Used During the Design Thinking Process? [online] Make:Iterate. Available at: https://makeiterate.com/why-are-personas-used-during-the-design-thinking-process/. Available at: (Accessed: .

[39] Oatley, K., Parrott, W.G., Smith, C. and Watts, F. (2011) 'Cognition and Emotion over twenty-five years', Cognition and emotion, 25(8), pp. 1341-1348 Available at: 10.1080/02699931.2011.622949.

[40] Obsidian Odyssey. (2023) 'Growth | Flow: The Psychology of Optimal Experience - Mihaly Csikszentmihalyi (1990)'.

[41] Östlund, U., Kidd, L., Wengström, Y. and Rowa-Dewar, N. (2011) 'Combining qualitative and quantitative research within mixed method research designs: a methodological review', International journal of nursing studies, 48(3), pp. 369-383.

[42] Park, Y.S., Konge, L. and Artino Jr, A.R. (2020) 'The positivism paradigm of research', Academic medicine, 95(5), pp. 690-694.

[43] Plass, J.L. and Kaplan, U. (2015) 'Emotional Design in Digital Media for Learning'Emotions, Technology, Design, and Learning, pp. 131-161.

[44] Prefit, A., Candea, D.M. and Szentagotai-Tătar, A. (2019) 'Emotion regulation across eating pathology: A meta-analysis', Appetite, 143, pp. 104438.

[45] Schneider, S., Nebel, S. and Rey, G.D. (2016) 'Decorative pictures and emotional design in multimedia learning', Learning and Instruction, 44, pp. 65-73.

[46] Stevens, E. (2019) 'How To Define A Problem Statement: Your Guide To The Second Step In The Design Thinking Process. [online] careerfoundry.com. Available at: https://careerfoundry.com/en/blog/ux-design/stage-two-design-thinking-define-the-problem/.', .

[47] Takahashi, F. and Kawabata, Y. (2018) 'The association between colors and emotions for emotional words and facial expressions', Color Research & Application, 43(2), pp. 247-257.

[48] Terninko, J. (2018) Step-by-step QFD: customer-driven product design. Routledge.

[49] V. Kompaniets, A. Lyz and A. Kazanskaya (2020) An Empirical Study of Goal Setting in UX/UI-design. pp. 1.

[50] Van Den Broek, E.L., Schut, M.H., Westerink, J.H., van Herk, J. and Tuinenbreijer, K. (2006) Computing emotion awareness through facial electromyography. Springer, pp. 52.

[51] Varghese, E., Jaggi, S., Gills, R. and Jayasankar, J. (2023) 'IBM SPSS Statistics: An overview', .

[52] Verduyn, P. and Brans, K. (2012) 'The relationship between extraversion, neuroticism and aspects of trait

affect', Personality and Individual Differences, 52(6), pp. 664-669 Available at: 10.1016/j.paid.2011.12.017.

[53] Vivek, R. and Nanthagopan, Y. (2021) 'Review and comparison of multi-method and mixed method application in research studies', European Journal of Management Issues, 29(4), pp. 200-208.

[54] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W. and Zhang, W. (2022) 'A systematic review on affective computing: emotion models, databases, and recent advances', Information Fusion, 83-84, pp. 19-52 Available at: 10.1016/j.inffus.2022.03.009.

[55] Whiteford, K.L., B. Schloss, K., E. Helwig, N. and E. Palmer, S. (2018) ' Color, Music, and Emotion: Bach to the Blues. ', (SAGE Journals) Available at: https://doi.org/10.25384/SAGE.c.4301789.v1.

[56] Winit-Watjana, W. (2016) 'Research philosophy in pharmacy practice: necessity and relevance', International Journal of Pharmacy Practice, 24(6), pp. 428-436.

[57] Xiao, Y. and Watson, M. (2019) 'Guidance on conducting a systematic literature review', Journal of planning education and research, 39(1), pp. 93-112.

[58] Yoon, J., Pohlmeyer, A.E., Desmet, P.M.A. and Kim, C. (2021) 'Designing for Positive Emotions: Issues and Emerging Research Directions', The Design journal, 24(2), pp. 167-187 Available at: 10.1080/14606925.2020.1845434.

[59] Yoshii, A., Plaut, D.A., McGraw, K.A., Anderson, M.J. and Wellik, K.E. (2009) 'Analysis of the reporting of search strategies in Cochrane systematic reviews', Journal of the Medical Library Association: JMLA, 97(1), pp. 21.

[60] Zhou, F., Ji, Y. and Jiao, R.J. (2020) 'Emotional Design', arXiv preprint arXiv:2010.03046, .

Innovations and Frontiers of Diffusion Models in Natural Language Processing: A Review

Jufang Zhao[1], Zengye Su[1*], Yudan Nie[1]
[1]School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou 511363, China
*Corresponding author: szy@xs.gcc.edu.cn

Abstract

Generative AI (GenAI) has emerged as one of the most transformative forces in artificial intelligence, profoundly impacting content creation, scientific research, and numerous application domains [1, 2]. At its core, these models learn the underlying distributions from existing data to generate novel, high-quality synthetic data [3]. Within this landscape, Foundation Models play a pivotal role. These are typically large-scale models pre-trained on massive datasets, possessing powerful generalization capabilities that serve as a robust baseline for various downstream tasks, thereby significantly reducing the development cost and time for AI applications [1]. Natural Language Processing (NLP), one of the first fields where GenAI achieved major breakthroughs, has largely benefited from the development of the Transformer model [4]. Since its introduction in 2017, the attention-based Transformer architecture has demonstrated outstanding performance on tasks such as machine translation, language understanding, and text generation. This has led to the development of foundation models, particularly Pre-trained Language Models (PLMs), which have greatly enhanced the performance of text generation tasks [5]. However, traditional text generation methods, especially autoregressive (AR) models, suffer from low inference efficiency when processing long texts [5, 6]. This paper provides a comprehensive review of Diffusion Models in NLP, exploring their fundamental principles, applications, and future directions.

Index Terms— Diffusion Models, Natural Language Processing (NLP), Generative AI, Text Generation, Transformer Models, Deep Generative Models, Literature Review.

1 Introduction

In recent years, diffusion models have emerged as a novel class of deep generative models, initially achieving breakthrough success in the image generation domain. They have surpassed previous state-of-the-art (SOTA) models, such as generative adversarial networks (GANs), in their ability to generate high-fidelity and diverse samples [7–10]. The core principle of diffusion models involves a forward process that systematically adds noise to data until it conforms to a simple prior distribution (e.g., random noise), followed by a learned reverse process that gradually denoises the signal to recover a data sample [3, 11, 12].

As research has progressed, the powerful capabilities of diffusion models have been explored for applications in Natural Language Processing, where they have shown immense potential in tasks like text generation [7, 13, 14]. Compared to traditional AR models and other generative frameworks like variational autoencoders (VAEs), GANs, and normalizing flows (NFs), diffusion models exhibit several distinct advantages in NLP [5, 6, 11]. Specifically, diffusion models offer greater flexibility in handling complex conditioning, as they can iteratively refine intermediate outputs based on given inputs to more easily generate high-quality target text [5, 6]. They also demonstrate inherent capabilities for global planning and self-correction, which are crucial for generating coherent and accurate long texts [15]. Furthermore, the training of diffusion models is generally more stable than that of GANs [11]. Although vanilla diffusion models can be slow at sampling, appropriate acceleration methods allow for an effective trade-off between inference time and generation quality [5, 6]. The ascent of diffusion models has even begun to challenge the long-held view that large language models must rely on the autoregressive paradigm, as illustrated in Figure 1, suggesting that the principles of generative modeling may be the true key to language intelligence [16].

Given the rapid development of diffusion models in NLP and their unique advantages, a systematic review of current research progress holds significant academic value and offers practical insights. While some surveys on diffusion models exist, they typically cover foundational principles, algorithmic variants, or applications in specific domains. A comprehensive review focused specifically on the application of diffusion models in NLP—particularly an in-depth analysis of model architectures, training methods, diverse application scenarios, and outstanding challenges—is currently lacking [9, 13].

This review aims to fill this gap by providing researchers and practitioners with a clear, comprehensive guide to the innovations and frontiers of diffusion models in the NLP domain. We systematically survey the development history, core model architectures, and training methodologies of diffusion models in NLP. The review focuses on analyzing their application in specific NLP tasks, including text generation (both autoregressive and non-autoregressive), text editing, and cross-modal generation, while also discussing their advantages and

Figure 1: Conceptual Comparison of Text Generation Processes

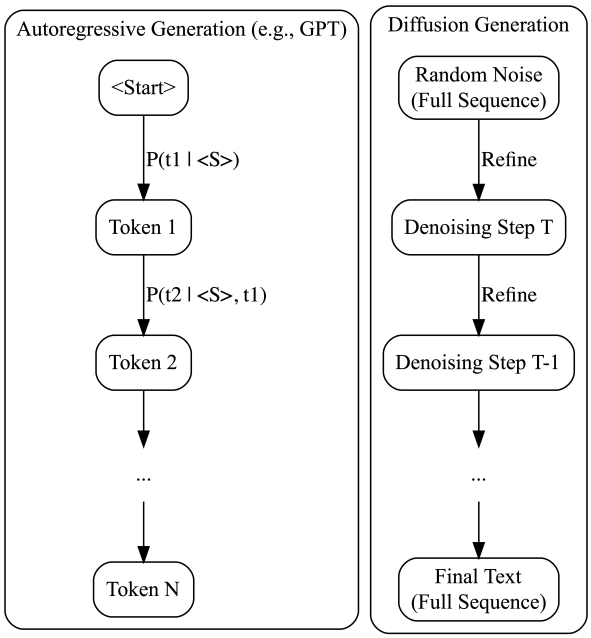


Figure 1: Conceptual comparison of text generation processes. **Left (Autoregressive):** Models like GPT generate text sequentially, predicting one token at a time. **Right (Diffusion):** Models start with random noise and iteratively refine the entire sequence in parallel.

limitations compared to traditional methods. Furthermore, we explore the integration of diffusion models with other prominent NLP models, such as Transformers and PLMs, and discuss current research challenges, including sampling efficiency, the modeling of discrete text data, and controllability. Finally, we provide an outlook on future research directions. By critically analyzing the contributions and shortcomings of existing work, this review seeks to highlight its novelty and value, thereby guiding future research and application of diffusion models in NLP.

The structure of this review is as follows: Section 2 details the foundational theory and principal variants of diffusion models. Section 3 focuses on the application of diffusion models in text generation and other NLP tasks. Section 4 discusses the integration of diffusion models with Transformer-based architectures. Section 5 covers optimization and acceleration techniques. Section 6 analyzes evaluation metrics and performance. Section 7 highlights the challenges and future outlook. Finally, Section 8 concludes the review.

2 Fundamentals of Diffusion Models

Diffusion models, an emerging class of generative models, draw inspiration from non-equilibrium thermodynamics to model complex data distributions by simulating a diffusion process [17]. (In this paper, we adopt the convention that bold lowercase letters, such as $\mathbf{x}$, denote vectors, while bold uppercase letters denote matrices). The core concept involves two processes: a forward process that progressively adds noise to an original data sample until it degenerates into a known prior distribution (typically a standard Gaussian), and a reverse process that learns to invert the forward process, gradually recovering a clean data sample from the noise [5,7,10,11,18,19]. This "corruption-and-reconstruction" paradigm provides a new framework for high-quality data generation [18].

Specifically, the forward diffusion process is modeled as a Markov chain of length T (timesteps) [5, 17]. In this process, the data $\mathbf{x}_0$ gradually evolves into noise $\mathbf{x}_T$. The transition probability $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ at each step, which describes the change from state $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$, is typically defined by the addition of Gaussian noise [11, 19]. By the Markov property, the joint probability of the entire forward process is given by Eq. (1):

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{1}$$

An important property of the forward process is that the noisy data $\mathbf{x}_t$ at any intermediate timestep t can be sampled directly from the original data $\mathbf{x}_0$. For Gaussian diffusion, this has a closed-form solution as shown in Eq. (2):

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_t = 1 - \beta_t$. The sequence of variances, $\{\beta_t\}_{t=1}^{T}$, is known as the noise schedule and is typically pre-defined to increase with t. As $t \to T$, $\bar{\alpha}_T$ approaches zero, such that $\mathbf{x}_T$ approximates a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ [11].

The reverse diffusion process aims to learn the inverse path, starting from the noise distribution $p(\mathbf{x}_T) = \mathcal{N}(0, \mathbf{I})$ and progressively denoising it to generate a data sample [11, 15]. This process is also modeled as a Markov chain, where the transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is approximated by a parameterized model, typically a deep neural network (the denoising network) [5, 11]. The training objective is to enable the learned reverse process to effectively recover the original data distribution from noise. This is typically achieved by maximizing the variational lower bound (VLB) on the data log-likelihood [16, 20]. In practice, the training objective is often simplified to minimizing the mean squared error (MSE) between the true added noise $\boldsymbol{\epsilon}$ and the noise predicted by the model $\boldsymbol{\epsilon}_\theta$, as shown in Eq. (3):

$$L_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2\right] \tag{3}$$

This loss function trains the model $\boldsymbol{\epsilon}_\theta$ to predict the noise that was added to the original data $\mathbf{x}_0$ to produce the noisy sample $\mathbf{x}_t$.

Since the proposal of Denoising Diffusion Probabilistic Models (DDPM) [19], diffusion models have garnered widespread attention. DDPM and its variants are among the most widely used diffusion frameworks today [14, 21]. In parallel, Score-Based Generative Models (SGM) and works that

unify both approaches within a Stochastic Differential Equation (SDE) framework have provided alternative mathematical perspectives on the data perturbation and recovery process [7, 20, 22].

2.1 Principles and Variants of Diffusion Models

Diffusion Models model complex probability distributions by simulating a dual-stage process: a Forward Diffusion Process and a Reverse Diffusion Process [3, 7, 12, 23].

In the forward process, a data sample $\mathbf{x}_0$ is progressively perturbed by adding noise over T timesteps, eventually transforming it into pure noise [5, 6, 17, 19]. This is modeled as a Markov chain, where the transition probability is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{4}$$

The reverse diffusion process aims to generate new samples by starting from pure noise and progressively denoising it [5, 17, 19]. This process is also modeled as a Markov chain, implemented via a parameterized reverse transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which is also modeled as a Gaussian:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{5}$$

where the mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\Sigma}_\theta$ are parameterized by a neural network, which typically adopts a U-Net or Transformer architecture [5, 6, 8, 11].

Variants of diffusion models differ in their principles and implementation. Denoising Diffusion Probabilistic Models (DDPM) are the canonical implementation [14, 18, 21]. Score-Based Generative Models (SGM) learn the data distribution's score function [7, 23]. Latent Diffusion Models (LDM) represent a significant advance in computational efficiency by operating in a compressed latent space [10, 18, 21, 24]. Other notable variants include Denoising Diffusion Implicit Models (DDIM), which accelerate sampling by defining a non-Markovian forward process [25].

2.2 Handling of Text Data

A core challenge in applying diffusion models to text generation is bridging the gap between the continuous-space formulation of diffusion and the discrete nature of text data [14]. Researchers have primarily pursued two categories of approaches: discrete text diffusion models and continuous text diffusion models [5, 13, 14].

The fundamental principle of a text diffusion model involves recovering a target text from a noisy input through a progressive denoising process [5, 6]. The reverse process can be generally expressed as:

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{c}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}) \tag{6}$$

where $\mathbf{c}$ represents the input condition [5, 6].

Discrete text diffusion models operate directly at the token level, generalizing the diffusion process to a discrete state space [14]. Continuous text diffusion models, conversely, encode discrete text into a continuous space where diffusion and denoising are performed [14]. Each approach has trade-offs regarding faithfulness to the data versus training stability and semantic richness [13].

2.3 Key Designs in the Diffusion Process

The performance of text diffusion models is critically influenced by four key design components: the denoising network, the noise schedule, the objective function, and the conditioning strategy [5, 6].

Denoising Network: The denoising network is the core of the reverse process. For sequential data like text, the Transformer architecture is widely adopted, as it allows the model to capture complex, long-range dependencies between tokens during the iterative refinement process [24].

Noise Schedule: The noise schedule defines the magnitude of noise added at each forward diffusion step. A well-designed schedule (e.g., linear or cosine) is crucial for generation quality, as it controls how quickly the original data signal is corrupted [19, 21].

Training Objective: The training objective is to learn the reverse denoising process, typically by minimizing the MSE between the predicted noise and the actual added noise (Eq. (3)). This simplified objective has been shown to be effective and stable for training high-quality generative models.

Conditioning Strategies: Conditioning strategies incorporate external information to guide generation. A powerful and common method is classifier-free guidance, which trains a single model to handle both conditional and unconditional generation, enabling strong, steerable synthesis at inference time [**?**, 5, 11].

3 Applications of Diffusion Models in NLP

Diffusion Models have achieved remarkable success in domains like image synthesis and are now showing significant potential to advance NLP tasks [26]. This section reviews their applications across NLP, detailing implementation methods, performance, and prospects.

3.1 Text Generation

Text generation aims to produce high-quality, coherent, and meaningful text. While traditional methods have been dominated by autoregressive models, diffusion models have recently emerged as a powerful alternative [12, 13]. Diffusion models offer unique advantages, including a natural fit for non-autoregressive (NAR) generation, better controllability, and flexible speed–quality trade-offs.

A variety of diffusion model variants have been developed for text generation. DIFFUSEQ and DIFFUSUM applied conditional diffusion to sequence-to-sequence tasks [13]. DIFFORMER, a Transformer-based model, showed strong perfor-

mance in machine translation and summarization [13]. The masked diffusion language model framework achieved state-of-the-art perplexity scores, demonstrating powerful modeling capabilities [27].

A key evaluation is comparison with autoregressive models like the GPT series. While GPT excels at fluent text generation [1], diffusion models offer complementary strengths. LLaDA, the first 8B-parameter diffusion-based large language model, demonstrates competitive performance with strong LLMs like LLaMA-7B in in-context learning [12, 16]. Notably, LLaDA mitigates the "reversal curse" seen in some AR models. However, limitations remain: some models are restricted to fixed-length text [13], and scaling diffusion models incurs significant computational cost [16].

3.2 Text Editing and Manipulation

Diffusion models are expanding into text editing and manipulation. Unlike autoregressive models, diffusion models treat editing as iterative denoising. For example, DiffusER conceptualizes edit operations as a noising process reversed by a denoising model [13]. The SUNDAE model handles arbitrary infilling within a template, providing a flexible framework for text repair [13]. This non-AR nature is advantageous for edits requiring global context. While direct quantitative comparisons with baseline editors are limited, the ability of diffusion models to perform text-guided image editing is well established, showcasing nuanced, instruction-based manipulation [25, 28, 29].

3.3 Text Representation Learning

In complex cross-modal tasks (e.g., text-to-image generation), text representations must capture fine-grained details. For instance, the Swinv2-Imagen model enhances text understanding by extracting entity and relationship embeddings from scene graphs [8]. However, the specific advantages of using diffusion models for representation learning per se remain unclear from current literature. Future research is needed to clarify the potential of diffusion models in this domain.

3.4 Machine Translation

Diffusion models have been applied to machine translation with promising results [13]. Several variants demonstrate strong translation capabilities. For example, the diffusion-based LLaDA model can effectively translate between Chinese, English, and German [16]. Other models like CDCD and SUNDAE also report high performance [13]. While these studies indicate excellent performance, they provide few details on diffusion's advantages for very long or complex sentences. Moreover, the potential of diffusion models for low-resource languages is intriguing but not yet detailed in available sources [30].

3.5 Dialogue Generation

In dialogue generation, diffusion models offer notable advantages, particularly in maintaining context and generating diverse responses [13, 16]. They effectively integrate multi-turn dialogue history; for example, LLaDA accurately captures extended conversation context [16]. The Latent Diffusion Energy-Based Model (LDEBM) is one approach that addresses issues like mode collapse by combining diffusion with an energy-based model [13]. Additionally, integrating external knowledge can further improve dialogue relevance [12].

3.6 Complex Reasoning Tasks

Diffusion models are being applied to complex reasoning in NLP. A notable development is the Diffusion of Thought (DoT) method [15], which introduces a chain-of-thought style reasoning within the diffusion framework. DoT performs reasoning by refining a sequence of latent "thought" variables in parallel over multiple steps. This allows multi-step reasoning to diffuse in parallel, offering a novel approach to tasks requiring several logical steps. DoT has been applied successfully to tasks needing sophisticated math and logic reasoning, demonstrating a powerful and novel reasoning mechanism [15].

3.7 Other Applications and Cross-Modal Fusion

Diffusion models are being extended to drive innovation in NLP and cross-modal tasks. In historical language studies, Oracle Bone Script Decipher (OBSD) uses diffusion to interpret ancient scripts [31]. In computer vision, OVDiff uses a diffusion model for open-vocabulary semantic segmentation without task-specific training [32]. Multimodal Diffusion Models, often within Multimodal LLMs, aim to process and fuse different modalities [9, 33, 34]. A common architecture uses a Transformer to create shared embeddings, which then condition a diffusion model [10, 24]. The Transfusion model enables seamless integration of discrete and continuous modalities within a single model [28].

4 Integration with Transformer Models

The Transformer architecture, with its powerful self-attention mechanism, is dominant in NLP [4, 18]. Given Transformers' prowess in sequence modeling, integrating them with diffusion models promises enhanced performance on complex generative tasks [14].

Diffusion Transformers (DiTs) exemplify this integration, replacing the typical U-Net backbone in vision diffusion models with a Transformer [24, 25]. In the multimodal domain, combining Transformers and diffusion is especially powerful. Latent Diffusion Models also often use Transformers to encode conditioning information (like text) into latent embeddings fed into the U-Net via cross-attention [10].

Table 1: Overview of Representative Diffusion Models in NLP and Related Fields.

Model (Year)	Primary Task / Domain	Key Architectural Feature	Reported Advantage / Contribution
Diffusion-LM (2022)	Unconditional Text Generation	Transformer in continuous embedding space	First to show diffusion LMs can achieve strong perplexity.
Masked Diffusion LM (2024)	Language Modeling	Transformer on discrete tokens with masking	Achieved state-of-the-art perplexity among diffusion models.
Differmer (2023)	Machine Translation, Summarization	Transformer-based denoising backbone	Competitive BLEU/ROUGE scores vs. strong Transformer baselines.
LLaDA (8B) (2024)	Instruction Following, Dialogue	Large-scale (8B param) diffusion LM	On-par with LLaMA-3 8B on dialogue benchmarks; mitigates repetition.
DoT (2024)	Mathematical/Logical Reasoning	Diffusion with parallel "thought" vectors	Outperforms AR chain-of-thought on some reasoning benchmarks.
Stable Diffusion (LDM) (2022)	Text-to-Image Generation	Diffusion in a compressed latent space (U-Net)	High efficiency and quality, enabling widespread use.

Beyond direct use, researchers are improving the Transformer architecture for diffusion. The DiffTransformer proposes a "differential attention" to reduce attention noise [4]. Its differential attention operator is calculated as:

$$\begin{aligned}\text{DiffAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \Bigg(&\text{softmax}\frac{\mathbf{Q}_1\mathbf{K}_1^T}{\sqrt{d_k}} \\ &- \lambda \cdot \text{softmax}\frac{\mathbf{Q}_2\mathbf{K}_2^T}{\sqrt{d_k}}\Bigg)\mathbf{V}\end{aligned} \tag{7}$$

where query, key, and value projections are split into two groups, and λ is a learnable scalar.

Will Transformers replace diffusion models? Current consensus is they are complementary rather than replacements [24]. Many SOTA models (LDM, DiT) combine both effectively. However, some work explores pure-Transformer generation, such as Google's Muse, which operates on discrete tokens and achieves SOTA text-to-image results efficiently without continuous diffusion [35].

5 Optimization and Acceleration

Despite their outstanding generative performance, diffusion models face challenges in computational cost and slow sampling [7, 22, 26, 36].

5.1 Sampling Acceleration

Slow sampling is a major bottleneck. Key acceleration strategies include:

- **Discretization Optimization:** Improving numerical solvers for the diffusion SDE/ODE [7, 15].
- **Non-Markovian Sampling:** Relaxing the Markov assumption to allow larger reverse steps. Denoising Diffusion Implicit Models (DDIM) [25] can cut steps from 1000 to as few as 50.
- **Distillation:** Progressive distillation trains a student model to perform two denoising steps of a teacher in one step, recursively halving inference time [7].

Additionally, efficiency improves by performing diffusion in a compressed latent space (as in LDM) [10, 21].

5.2 Maximum Likelihood Estimation Enhancement

Improving log-likelihood is crucial [7]. Methods include:

- **Noise Schedule Optimization:** Nonlinear schedules like cosine can improve performance [19, 21].
- **Objective Design:** Tailoring the loss to the task can yield better results [5].
- **Learnable Reverse Variance:** Learning the reverse process variance can improve likelihoods [7].

5.3 Model Architecture and Inference Acceleration

Refining the network architecture is another key lever [21, 24]. For inference, model compression is widely used [7, 18]:

- **Knowledge Distillation:** A large teacher model trains a smaller student model to speed up inference [7, 25].
- **Pruning and Quantization:** Techniques like QLoRA combine 4-bit quantization with low-rank adaptation for efficient fine-tuning and inference [18].

6 Evaluation Metrics and Performance Analysis

Evaluating diffusion models requires diverse metrics to assess quality, fidelity, diversity, and efficiency. For cross-modal generation (e.g., text-to-image), standard metrics are Fréchet Inception Distance (FID) and CLIP score [8, 24, 28, 35, 37]. For text generation, traditional metrics include BLEU, ROUGE, and perplexity [37]. However, these often miss nuanced qualities like global coherence or logical consistency, indicating a need for more comprehensive evaluation protocols for diffusion-generated text [12]. For tasks requiring precise correctness (like reasoning), accuracy is key [15]. Several factors influence performance including model architecture, data scale/quality, and training strategy. Despite SOTA results, diffusion models have known challenges. They can be sensitive to input noise and remain computationally intensive [33]. More critically, current automatic metrics for text often fail to capture high-level attributes of generated text [12].

7 Challenges and Future Outlook

Despite their potential, diffusion models in NLP face several challenges [6]:

- **Computational Cost and Sampling Speed:** High cost and slow generation remain prominent issues [10, 12, 20, 22, 37].
- **Discrete Data Modeling:** A fundamental mismatch exists between discrete text and continuous diffusion formulations [5, 6].
- **Interpretability and Controllability:** Diffusion processes are less interpretable, and fine-grained control remains challenging [3, 20].
- **Data, Safety, and Bias:** Diffusion models can learn societal biases or produce harmful content. Developing methods for content moderation and "detoxification" is crucial for responsible AI deployment [1, 3, 5, 6].
- **Multilingual and Low-Resource Scenarios:** Extending diffusion models to multilingual or low-resource settings is largely unexplored and will require innovative strategies.

Future Outlook: The future of diffusion models in NLP is promising, with key directions including:

1. **More Powerful and Efficient Models:** Continue scaling model size and exploring novel architectures, while developing training and sampling methods that improve efficiency [8, 11, 25].
2. **Broader NLP Applications:** Apply diffusion models to a wider range of tasks, including analytical tasks, knowledge graph construction, and structured prediction [5, 7, 22].
3. **Synergy with Other Technologies:** Deeper integrate diffusion with PLMs, combine with Transformers and knowledge graphs, and develop unified multimodal models [5, 6, 28].
4. **Advancing Language Representation:** Move toward a continuous language space for representing text, eliminating discrete tokenization limits [12].
5. **Improved Evaluation and Responsible AI:** Create more holistic evaluation benchmarks and focus on reliability, controllability, and bias mitigation to ensure safe deployment [9, 34].

The rise of foundation models and generative AI will continue to shape diffusion models' trajectory in NLP [1, 2].

8 Conclusion

This review has provided a comprehensive overview of diffusion models in NLP. As powerful generative tools [3], diffusion models have shown a remarkable ability to generate high-quality data [23]. Their innovative role and immense potential in NLP are increasingly evident [14].

Current research demonstrates significant advantages on core NLP tasks like text generation and editing [14]. Compared to AR models, diffusion models excel in parallel generation and fine-grained controllability. Advanced applications such as Diffusion-of-Thought show potential to surpass the AR paradigm for complex reasoning tasks [15]. Large-scale models like LLaDA are now competitive with traditional LLMs, while offering unique benefits like bidirectional generation [16].

However, challenges remain, including high computational costs, difficulty modeling discrete text, and interpretability and safety issues [3]. Effectively modeling diffusion for text, optimizing sampling efficiency, and better leveraging PLMs are key open questions.

Looking ahead, the potential of diffusion models in NLP is vast. Central focus will be on more efficient training and inference [37]. Architectural innovation—particularly integrating Transformers [4]—will be critical. The synergy between these technologies will likely spur novel applications in low-resource languages, advanced text analysis, and multimodal fusion [11]. In summary, diffusion models are bringing new vitality to NLP. While challenges persist, ongoing research is poised to overcome these obstacles, fully unleashing their

power to build more intelligent and creative language technologies. Ultimately, by bridging the gap between parallel and sequential processing, diffusion models are not just a new tool for NLP but a step towards more flexible and human-like language intelligence.

Funding

This work is funded by the Guangdong Provincial Sci-Tech Innovation Strategy Fund [pdjh2024a467].

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, and others. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] I. Toumi, I. Rjiba, S. Ben Abdallah, and A. M. Alimi. Generative ai: systematic review of advancements. *Multimedia Tools and Applications*, pages 1–39, 2024.

[3] Chenshuang Zhang, Chaoning Zhang, Meng Zhang, and Hedvig Kjellstrom. Text-to-image diffusion models in generative ai: A survey. *IEEE Transactions on Multimedia*, 2023.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008, 2017.

[5] Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905, 2024.

[6] Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion models for non-autoregressive text generation: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6298–6306. ijcai.org, 2023.

[7] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[8] Jialu Sui, Jianzong Wang, Shijing Si, Zhangcheng Huang, and Jing Xiao. Swinv2-imagen: text-to-image generation via hierarchical models. *Neural Computing and Applications*, 36(13):11119–11129, 2024.

[9] Jian-Wei Zhang, Han-Jia Chen, Jian-She Tan, Run-Ze He, Kun Zhang, Tao Qin, and Tie-Yan Liu. A survey on controllable diffusion models. *Journal of Computer Science and Technology*, 39(4):923–955, 2024.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2022.

[11] Zhaoyu Chen, Yuerong Chen, Zijie Yue, Yihang Luo, Shanshan Li, Pedram Ghamisi, and Beichen Zhang. Diffusion models in remote sensing image processing: A review and outlook. *arXiv preprint arXiv:2404.08926*, 2024.

[12] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 3036–3053, 2022.

[13] Hao Zou, Zae Myung Kim, and Dongyeop Kang. A survey of diffusion models in natural language processing. *arXiv preprint arXiv:2305.14671*, 2023.

[14] Xiao Han, Sheng He, ZipeGASUS Li, and Stan Z. Li. A survey on diffusion models in nlp. *arXiv preprint arXiv:2305.14387*, 2023.

[15] Boming Pang, Chen Meng, Qing Han, and Kun He. Diffusion of thought: A new potential for llms. *arXiv preprint arXiv:2402.07754*, 2024.

[16] Hong-Yi Lin, Zhaowei Zhang, Chenghao Chi, Lichao Yu, Yunchao Wang, Haotian Liu, Chunyuan Wu, Peng Li, and Lijuan Wang. LLaDA: A Foundation Model for Bidirectional Language Generation. *arXiv preprint arXiv:2406.18349*, 2024.

[17] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015.

[18] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative models: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11674–11693, 2023.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 6840–6851, 2020.

[20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[21] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of*

the 38th International Conference on Machine Learning (ICML), pages 8162–8171, 2021.

[22] Bowen Jing, Morteza Eslami, Ezra Miller, Peter J.M. Claes, Tom Sercu, Alexander M. Rush, and Frederick P. Roth. Diffusion models are a new generation of generative ai for biology. *Nature Computational Science*, 3(11):923–933, 2023.

[23] Han Cao, Cheng Tan, Zhangyang Gao, Yitan Li, Siyuan Liu, Pin-Yu Xie, Li Zhang, Jian-Li Li, and Jian Gao. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205. IEEE, 2023.

[25] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image edit instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402. IEEE, 2023.

[26] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[27] Sheng He, Jiacheng Chen, Kang Zhou, and Zhe Zhao. Masked diffusion language models are latent variable models. *arXiv preprint arXiv:2406.11181*, 2024.

[28] Zhaowei Cai, Riza Velioglu, Yuxuan Fang, Bora Alten, Avinash Ravichandran, Anurag Arnab, Chen Sun, Ziad Al-Halah, and Stefano Soatto. Transfusion: A Unified Generative Model for Text, Image, and Multi-Modal Tasks. *arXiv preprint arXiv:2406.13110*, 2024.

[29] Michal Avrahami, Dani Lischinski, and Ohad Fried. Blended-latent-diffusion. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11. ACM, 2023.

[30] Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57, 2014.

[31] Zihan Li, Yixuan Gou, Xiang Fan, Zhiqiang Zuo, Peng Li, Ming-Hsuan Yang, Zhenglin Ye, and Cheng-Chuan Ping. Lost in translation: Reimagining the classic cipher and lost language problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1020–1034, 2024.

[32] Zeren Xu, Zhiyong Zhang, Jin Tang, and Chang Xu. Open-vocabulary semantic segmentation with diffusion models. *arXiv preprint arXiv:2307.03131*, 2023.

[33] Jiawei Zhang, Jialing Jia, Yuan Zhang, Lin Luo, Wei Wang, and En Zhang. VL-3DDet: A new paradigm for vision-language models in 3D object detection. *arXiv preprint arXiv:2405.18738*, 2024.

[34] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Zhan. A survey of multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[35] Huiwen Chang, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Gaddy, and others. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, Cham, 2015.

[37] Calvin Gao, Jiaming Li, Jun Zhu, and Maciej Bogun. T-MARS: A-Posteriori-Controllable Text-to-Image Generation. *arXiv preprint arXiv:2403.01824*, 2024.