INTERNATIONAL JOURNAL of Advanced AI Applications

Volume2,Issue6,Jun.2026

Online ISSN 3104-9338

Print ISSN 3104-932X

Hong Kong Dawn Clarity Press Limited

http://www.dawnclarity.press/index.php/ijaaa

TABLE OF CONTENTS

Online ISSN: 3104-9338
Print ISSN: 3104-932X

China-Japan Classic Literature Shared Reading Assisted by AI: A Case Study of The True Story of Ah Q and I Am a Cat

Huyue Mao*, Yanhua Xu

Zhejiang Yuexiu University

Received: May 7, 2026

Revised: May 8, 2026

Accepted: May 9, 2026

Published online: May 16, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 6 (Jun 2026)

* Corresponding Author: Huyue Mao (3400409664@qq.com)

Abstract. Against the backdrop of contemporary globalization, cross-cultural communication between China and Japan has become increasingly frequent. However, current Chinese-Japanese language teaching and learning predominantly focus on vocabulary and grammar, with insufficient attention paid to the cross-cultural analysis and exchange of classic literary works. Traditional cultural exchanges between the two countries have largely operated as one-way output, lacking genuine bilateral interaction. This project adopts a "shared reading" approach, in which Chinese and Japanese students read translated versions of each other's literary classics, thereby engaging in cross-cultural communication from the dual perspectives of readers and translators. Using Lu Xun's The True Story of Ah Q and Natsume Sōseki's I Am a Cat as examples, the project implements shared reading sessions and analyzes the two books' translations from three dimensions: language, culture, and style. The research reveals that while the translations exhibit certain problems — such as the loss of colloquial tone, simplification of culture-specific terms, and weakening of satirical style — both works nevertheless resonate with participants in their critique of human nature. The China-Japan shared reading model, to some extent, breaks through the limitations of traditional cultural exchange, fostering equal dialogue between the two countries' classic literary traditions. It can serve as a reference for cross-cultural activities in higher education.

Keywords: *China-Japan Shared Reading; The True Story of Ah Q; Cross-cultural Communication; Translation Comparison; AI*

1. Introduction

1.1. Research Background

With the continuous development of international relations and globalization, cultural exchanges between China and Japan have grown increasingly close. However, the current focus of exchange remains concentrated on popular culture fields such as anime, film, television dramas, and video games. The mutual reading and in-depth discussion of classic literary works remain relatively limited. Furthermore, Chinese-Japanese language teaching in both countries tends to emphasize vocabulary acquisition and grammatical accuracy, while much less attention is devoted to the exploration and analysis of each other's literary traditions. As a result, students often struggle to understand the deeper cultural connotations embedded in the language they are learning.

In addition, the promotion of Chinese and Japanese cultures has largely followed a one-way output model. China has translated a vast number of Japanese literary works, but these translations are primarily read by Chinese readers themselves, with few opportunities for dialogue with Japanese readers. Similarly, Japan has produced numerous translations of Chinese literary works, yet it also lacks a bilateral platform for exchange and discussion. This unidirectional model of translation and reading makes it extremely difficult to achieve genuine cross-cultural communication — a process that requires mutual understanding and dialogue rather than mere one-way transmission.

1.2. Research Methods

Against this background, the project titled "China-Japan Shared Reading: When Ah Q Speaks Japanese, when a Cat Speaks Chinese" uses two classic literary works as a bridge to build a platform for cross-cultural interaction among young students. The project organizes Chinese and Japanese students into shared reading groups, where they engage in in-depth reading of both books. Through this process, the participants analyze translation differences from three perspectives: language, culture, and style. By reading the Japanese translation of The True Story of Ah Q and the Chinese translation of I Am a Cat together, students from both countries are able to compare the original texts with their translated versions, discuss their observations, and develop a deeper understanding of each other's cultural and linguistic nuances.

This shared reading project is implemented with the assistance of Artificial Intelligence (AI).

First, when Chinese and Japanese students encounter words that are difficult to translate or comprehend during shared reading, they can use AI to obtain detailed explanations and

supplementary notes for these terms, which effectively removes their reading barriers.

Second, as Chinese and Japanese students lack familiarity with each other's cultural backgrounds, AI provides them with one-click access to historical contexts — including the 1911 Revolution, the Meiji Period, social customs and core values — helping students immerse themselves in the specific historical era and gain a deeper understanding of the literary characters.

Third, acting as an impartial intermediary free from cultural biases, AI can elaborate on the respective perspectives of both sides in their native languages, thereby reducing misunderstandings and cultural cognitive deviations arising from inadequate linguistic communication.

2. The Basic Framework of the China-Japan Shared Reading Model

2.1. Theoretical Foundations

This project draws upon several key theoretical frameworks.

First, cross-cultural communication theory suggests that genuine cross-cultural understanding does not merely involve accepting another culture on its own terms. Rather, it requires the ability to engage in dialogue from both perspectives simultaneously. This means that participants must be willing to see themselves as both learners and teachers, constantly moving between their own cultural framework and that of the other.

Second, the concepts of "domestication" and "foreignization" in translation studies provide useful analytical tools for examining translation differences. Domestication refers to translation strategies that make the target text read fluently and naturally, as if it were originally written in the target language. Foreignization, by contrast, refers to strategies that deliberately preserve elements of the source text that may seem strange or unfamiliar to target readers, thereby retaining the "foreignness" of the original. Both strategies have their respective advantages and disadvantages, and their use often reflects deeper cultural assumptions.

Third, reader-response theory posits that the same text can generate significantly different interpretations among readers from different cultural backgrounds. According to this theory, meaning is not fixed within the text itself but is actively constructed by readers based on their own experiences, cultural knowledge, and interpretive frameworks. These three theoretical perspectives — cross-cultural communication theory, domestication/foreignization in translation studies, and reader-response theory — together provide a solid foundation for the

shared reading model.

2.2. The Core Logic of the Shared Reading Model

When Chinese and Japanese students read the Japanese translation of The True Story of Ah Q together, several processes occur simultaneously. For Chinese students, reading the Japanese translation not only improves their Japanese reading ability but also allows them to discover how a classic Chinese literary character — Ah Q — is perceived and presented by Japanese readers and translators. They can observe the subtle differences between the original Chinese Ah Q and the "Japanese" Ah Q, differences that often reveal much about how Japanese culture interprets Chinese literature.

For Japanese students, the process is equally valuable. By helping their Chinese peers understand the Japanese translation of The True Story of Ah Q, Japanese students can better grasp the ideas and cultural emotions expressed in this foundational Chinese literary work. They gain insights into how Chinese readers might interpret their own explanations and how translation choices shape cross-cultural understanding.

Similarly, when reading the Chinese translation of I Am a Cat together, Chinese students assist Japanese students with Chinese reading comprehension while simultaneously absorbing the thoughts and emotions expressed in this classic Japanese work. Japanese students, in turn, discover differences between the original Japanese text and its Chinese translation, thereby experiencing firsthand how Chinese culture interprets and reinterprets Japanese literature. Through this shared reading model, participants not only enhance their foreign language abilities but also develop a deeper appreciation for the differences between original works and their translations. The reading process becomes an active exploration of cultural differences and shared understandings, promoting language acquisition and cross-cultural communication simultaneously.

3. Translation Comparison of The True Story of Ah Q and I Am a Cat

The translation comparison of the two works is organized around three analytical dimensions: language, culture, and style.

3.1. Linguistic Differences

The True Story of Ah Q makes extensive use of colloquial, vernacular, and streetwise satirical expressions. Lu Xun deliberately employed plain, unadorned language to depict the life and

psychology of a lower-class individual in early 20th-century rural China. For example, the original Chinese phrase "确乎很值得惊异" carries a deliberate, measured, almost leisurely tone. Its Japanese translation, 「実に驚くべきことだった」, converts this into standard written language, thereby weakening the original's subtle satirical edge.

Another telling example is the coarse expression "妈妈的" (a mild minced oath in Chinese, literally something like "mother's"). This expression, which conveys anger but not extreme profanity, carries the flavor of a Chinese lower-class insult. The Japanese translation renders it as 「ちくしょう」, which roughly means "beast" or is used as an expletive roughly equivalent to "damn it." While the Japanese translation retains the sense of anger, it loses the vernacular, folk quality of the original — a quality that reveals much about Chinese rural speech patterns. This example reflects the Japanese translation's tendency to prioritize semantic equivalence (conveying what words mean) while sacrificing stylistic register (how those words sound in context).

In I Am a Cat, the original Japanese text possesses strong classical colloquial features. For instance, the expression 「とんと見当がつかぬ」 (completely without any idea) and the pronoun 「吾輩」 (wagahai, an archaic, self-important way of saying "I" or "one") both evoke a particular period and style of Japanese writing. The Chinese translation renders 「とんと見当がつかぬ」 as "压根儿不知道" (yāgēnr bù zhīdào, meaning "not know at all," with a colloquial northern Chinese flavor) and 「吾輩」 as "我" (standard "I"). This relatively straightforward tone effectively conveys the casual, self-mocking quality of the cat's narration. This comparison indicates that the Chinese translation places greater emphasis on stylistic register — on preserving the way the narrative sounds — compared to the Japanese translation of The True Story of Ah Q.

3.2. Cultural Differences

In The True Story of Ah Q, terms such as "Weizhuang" (the name of Ah Q's village), "xiucai" (a degree holder under the imperial examination system), and "foreign devil" (a derogatory term for Westerners) all carry specific historical and cultural meanings that emerged in modern Chinese history. "Wei Zhuang" represents the closed, hierarchical, and tradition-bound nature of rural China during the period of the 1911 Revolution. "Xiucai" refers to individuals who had passed one level of the imperial examinations — a system that had shaped Chinese society for over a thousand years but was abolished in 1905. "Foreign devil" reflects the xenophobic sentiments prevalent among ordinary Chinese people during the late Qing dynasty and early

Republican period.

The Japanese translation renders 「秀才」 as 「紳士」 (shinshi, meaning "gentleman"). This is a clear example of domestication: the translator uses a term familiar to Japanese readers to replace one that would otherwise require lengthy explanation. While this domestication strategy certainly improves readability for Japanese readers, it also detaches the term from its historical and cultural context. As a result, Japanese readers may struggle to understand Ah Q's ambivalent attitude toward the xiucai — an attitude that mixes contempt for those who passed the exams with envy of their social status.

By contrast, the Chinese translation of I Am a Cat directly retains culturally specific Japanese terms such as "tatami" (traditional straw mats), "shoin" (a study or writing room), and "shōji" (paper sliding doors). The translation does not domesticate these terms into Chinese equivalents such as "floor mats" or "study." This foreignization strategy preserves the distinctiveness of Japanese culture. Moreover, because many Chinese readers already have some familiarity with these terms through exposure to anime, Japanese television dramas, and other forms of popular culture, the reading threshold is not significantly raised. Thus, the Chinese translation achieves effective cultural transmission without sacrificing readability.

3.3. Stylistic Differences

Lu Xun is known for his stark, understated style — sometimes called "cold white drawing" (冷峻白描) — a technique through which he criticized what he saw as the weaknesses of the Chinese national character. For example, in The True Story of Ah Q, Lu Xun writes: "He felt that since the world began, some people were bound to be arrested and brought to court, and some were bound to draw circles on a piece of paper — it was nothing to take to heart." This passage, on its surface, appears to show Ah Q consoling himself after a humiliating experience. But beneath the surface, it reveals his profound numbness: he has internalized his powerlessness to such an extent that even his own degradation no longer disturbs him. The Japanese translation transforms this cold, sharp satire into a more moderate, less pointed narrative, thereby somewhat weakening the original's critical force.

Natsume Sōseki, a central figure in the Yoyū-ha (School of Detachment), employed a style of satire that is characteristically understated, indirect, and subtle. His criticisms are delivered with a light touch, often through the voice of an ironic observer. In the Chinese translation of I Am a Cat, some of Sōseki's more implicit expressions are rendered more directly. For instance, phrases such as "任性又麻木" (selfish and numb) and "空谈一事无成" (empty talk, nothing

accomplished) convey the critical sentiment clearly and forcefully. This directness increases cross-cultural accessibility: Chinese readers can easily grasp what Sōseki is criticizing. At the same time, however, something of the original's "detached" quality — the subtlety and lightness that characterizes Sōseki's humor — is inevitably lost in the process.

3.4. Points of Resonance Between the Two Works

ls during periods of rapid social transformation. Ah Q's famous "spiritual victory" (精神胜利法) — his habit of mentally claiming victory to mask real-world failure — is one such coping mechanism. Kuzushū (the master in I Am a Cat), who engages in endless empty talk and pedantic discourse while avoiding meaningful action, represents another. Both are weak individuals who, faced with sweeping changes that they cannot control, develop pathological ways of coping that only deepen their powerlessness.

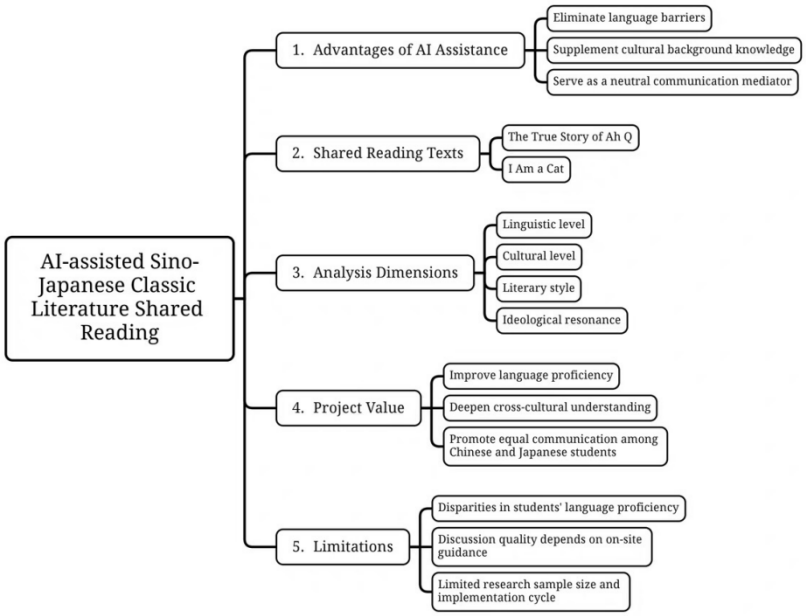

Figure 1. Mind map of the shared reading project.

During the shared reading sessions, both Chinese and Japanese students reported finding points of resonance in these themes. Some Japanese students felt that Ah Q's spiritual victory reminded them of the psychological coping strategies observed among the "lost generation" of Japanese youth who came of age during the prolonged economic stagnation following Japan's asset price bubble burst in the early 1990s. Some Chinese students noted that Kuzushū's state of "knowing a lot but doing nothing" can be observed among many young people in contemporary China, especially those facing intense competitive pressure in education and employment. These observations demonstrate the cross-cultural value of classic literature: the

"human weaknesses" they portray are not peculiar to any single national character but represent universal commonalities that emerge in societies undergoing modernization.

4. Practical Effectiveness and Limitations of the Shared Reading Project

4.1. Practical Effectiveness

The project produced three main areas of practical benefit.

First, differences between the Chinese and Japanese translations stimulated rich, in-depth discussion. When participants encountered the translation of "xiucai" as 「紳士」 (gentleman), Chinese students explained to their Japanese peers the historical nature of the imperial examination system and its profound social significance in pre-modern China. In return, Japanese students explained the semantic differences between the Japanese expression 「気味が悪い」 (kimochi ga warui, an expression of unease or creepiness) and the Chinese word "恶心" (ěxīn, meaning "nauseating" or "disgusting"). This mutual process of explanation was something that neither group could have achieved by reading alone.

Second, the shared themes identified in both works fostered genuine emotional connections among participants. The resonance that Chinese and Japanese students both felt regarding spiritual victory and empty pedantry allowed their discussions to move beyond mere textual analysis and enter the realm of personal reflection and cross-cultural empathy. Rather than impeding understanding, cultural differences in interpretation became the starting point for deeper mutual comprehension.

Third, the shared reading experience enhanced participants' language skills and cultural sensitivity. By comparing translations side by side, participants noticed details that they typically overlooked when reading alone. They learned to use translation discrepancies as entry points for exploring deeper cultural questions — a skill that has lasting value beyond the project itself.

4.2. Limitations

The shared reading model also presents certain limitations. The depth and quality of the discussion are constrained by participants' proficiency in the foreign language as well as their level of prior knowledge regarding each other's national histories. Participants with limited language ability struggle to perform meaningful translation comparisons. Participants who lack basic knowledge of the 1911 Revolution or the Meiji Restoration find it difficult to situate

literary characters within their specific historical contexts. Furthermore, the model requires a certain degree of facilitation skill on the part of discussion leaders. Without skilled facilitation, discussions risk remaining superficial or veering off-topic. Finally, because this project involved a relatively small participant sample and a limited implementation period, the generalizability of its conclusions — the extent to which they can be applied to other settings and other literary works — requires further testing.

5. Conclusion

This project used translation comparisons of The True Story of Ah Q and I Am a Cat as its entry point. By analyzing the two works from the perspectives of language, culture, and style, the research revealed that the translations exhibit varying degrees of loss and adjustment in conveying the original texts' colloquial tone, culture-specific terms, and satirical style. The Japanese translation of The True Story of Ah Q tends toward domestication and stylistic softening. The Chinese translation of I Am a Cat comparatively favors foreignization and more direct expression. These differences are likely inevitable outcomes of the different cultural aesthetics, translation traditions, and target reader expectations that shape translation choices in China and Japan.

Despite these differences, both works' shared critique of human weaknesses — their unflinching portrayal of how weak individuals cope with powerlessness and social change — produced sufficient cross-cultural resonance to support meaningful shared reading practice. Both Chinese and Japanese participants found something recognizable in Ah Q's spiritual victory and Kuzushū's empty pedantry.

The practice of the shared reading model demonstrates its effectiveness in breaking through the limitations of one-way cultural output. By engaging in translation comparison and mutual explanation, learners can enhance both their language abilities and their cultural understanding in a process that is simultaneously collaborative and reflective. Moreover, the shared reading model incurs low implementation costs, does not require sophisticated technology, and exhibits strong replicability. It can therefore be promoted and adapted for use in university-level language learning contexts beyond the two books examined here.

The primary limitations of this study are its small sample size and its brief, single-semester implementation period. Future research could expand the range of works examined — perhaps including poetry, drama, or modern prose — to further test the model's effectiveness and generalizability.

References

[1] Lu Xun. (2020). 阿Q正传 [The True Story of Ah Q]. Y*angtze River Literature and Art Publishing House.* [in Chinese]

[2] Lu Xun. (2020). 阿 Q 正传（日文版） （井上红梅， 译）[The True Story of Ah Q (Japanese Edition) (Translated by Inoue Kōbai)]. *East China University of Science and Technology Press.* [in Chinese]

[3] Natsume Sōseki. (2020). 我是猫 （林少华， 译）[I Am a Cat (Translated by Lin Shaohua)]. *Qingdao Publishing House*. [in Chinese]

[4] Natsume Sōseki. (2020). 我是猫. [I Am a Cat]. *East China University of Science and Technology Press*. [in Chinese]

[5] Sun Qiaona. (2014). 《我是猫》与《阿 Q 正传》中的社会批评比较研究 [A Comparative Study of Social Criticism in 'I Am a Cat' and 'The True Story of Ah Q']. *China National Knowledge Infrastructure*. [in Chinese]

[6] Xie Nandou. (2003). 从《我是猫》到《阿Q正传》 [From 'I Am a Cat' to 'The True Story of Ah Q']. *Chinese Literature Studies*, (1), 61–63. [in Chinese]

[7] Hu Min. (April 28, 2026). 成都市“AI+教育”迈向规模化应用 [Chengdu's 'AI Education' is moving towards large-scale application]. *Sichuan Economic Daily*, 003. [in Chinese]

[8] Shang Qiaoqiao, Wang Jingying, & Lun Yingjie. (2021). 国外教育机器人辅助语言学习研究：认知激活与情感驱动 [Research on Foreign Educational Robots Assisting Language Learning: Cognitive Activation and Emotional Drive.]. *Digital Education*, 7(1), 79–84. [in Chinese]

[9] Wang Dongfeng. (2025). AI 时代的文学翻译 [Literary Translation in the AI Era.]. *Foreign Language Audio-Visual Teaching*, (6), 3–11. [in Chinese]

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Intelligent Project Reimbursement Material Organization Mini-Program based on OCR Technology

Xinrui Liu

Chengdu University of Information Technology, Chengdu, China

Received: May 28, 2026

Revised: May 29, 2026

Accepted: May 30, 2026

Published online: May 31, 2025

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 6 (June 2026)

* Corresponding Author: Xinrui Liu (3171788575@qq.com)

Abstract. The increasing volume and complexity of project reimbursement materials in universities have highlighted the limitations of traditional manual processing methods, which are time-consuming, error-prone, and inefficient. This paper presents an intelligent mini-program based on Optical Character Recognition (OCR) technology designed to automate the organization and data extraction of reimbursement documents. The system integrates a WeChat mini-program front-end with a Flask back-end server and a MySQL database, employing a multi-strategy OCR scheduling mechanism that dynamically selects the most appropriate OCR engine (e.g., Tesseract for printed text, commercial engines for handwritten or low-quality images) based on document characteristics. Extracted data are cleaned and validated using regular expressions, and a finite state machine manages real-time processing status tracking. The system also generates standardized Excel reports compatible with university financial department requirements. Experimental results demonstrate that the proposed system achieves 97.3% character recognition accuracy and 94.7% field extraction accuracy, with an average response time of 3.2 seconds per document, outperforming Tesseract OCR and Abbyy FlexiCapture baselines. The mini-program significantly improves efficiency and accuracy in reimbursement material processing, reduces manual labor, and promotes digital transformation in university financial management. This research contributes to the application of AI-driven document intelligence in industry, providing a scalable and adaptable solution for automated financial workflows.

Keywords: *OCR Technology; Intelligent Mini-Program; Project Reimbursement; Edge AI; AI in Industry*

1. Introduction

1.1. Research Background

In the context of the rapid development of information technology, universities have witnessed a significant increase in scientific research projects, accompanied by a corresponding growth in the volume and complexity of project reimbursement materials. Traditional reimbursement processes often require researchers to manually prepare and organize a large number of paper-based documents, including invoices, contracts, and expense receipts, which are not only time-consuming but also error-prone. Furthermore, the current trend towards digital transformation has accelerated the adoption of electronic invoices; however, many universities still rely on legacy systems that treat electronic invoices similarly to paper invoices, resulting in inefficient reimbursement workflows. To address these challenges, there is an emerging demand for intelligent solutions that can leverage advanced technologies such as Optical Character Recognition (OCR) to automate and streamline the reimbursement material organization process. OCR technology, with its ability to convert scanned or image-based documents into machine-readable text, holds great potential for enhancing the efficiency and accuracy of financial management in universities [1, 2].

1.2. Problem Statement

The traditional approach to organizing project reimbursement materials faces several critical issues that significantly hinder the overall efficiency of financial management in universities. First and foremost, the manual entry of data from both physical documents and electronic files is highly prone to errors, ultimately leading to delays in reimbursement approval and potential financial losses for the institution. Furthermore, the complex and diverse formats of reimbursement materials—including handwritten notes, low-quality scanned images, and multi-column tables—pose significant challenges for accurate data extraction and processing. Additionally, the large volume of materials generated by numerous scientific research projects results in a particularly heavy workload for financial personnel, who must spend considerable time verifying, organizing, and cross-referencing these documents manually. These persistent problems not only increase labor costs substantially but also impede the overall digitalization process of university financial management systems. Therefore, developing an intelligent automated system that can efficiently extract, clean, and organize reimbursement data is imperative to alleviate these pressing issues and improve both the efficiency and accuracy of the reimbursement process considerably.

1.3. Research Objectives

This research aims to develop an intelligent mini-program based on OCR technology to address the challenges associated with project reimbursement material organization in universities. Specifically, the system is designed to achieve the following objectives:

(1) Improve the efficiency of reimbursement material processing by automating the extraction of key information from various document types;

(2) Enhance data accuracy through advanced OCR scheduling strategies and regular expression-based data cleaning methods;

(3) Provide a user-friendly interface to facilitate real-time interaction and status tracking throughout the processing workflow.

By integrating OCR technology with other advanced technologies such as Flask backend development, MySQL database management, and JWT authentication, the proposed system is expected to make significant contributions to the field of AI in industry, particularly in the area of financial automation and document intelligence processing. Moreover, the successful implementation of this system is expected to promote the digital transformation of university financial management and serve as a reference for intelligent solutions in other industries and scenarios.

2. Related Work

2.1. OCR Technology Development

Optical Character Recognition (OCR) technology has evolved significantly since its inception, from simple pattern matching algorithms to complex deep learning models [1]. The early stages of OCR development focused on rule-based approaches that relied on template matching and feature extraction techniques. These methods were effective for recognizing printed characters with limited font variations but struggled with handwritten text or complex backgrounds. With the advancement of machine learning, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were introduced to improve the accuracy and robustness of OCR systems. In recent years, deep learning has revolutionized the field by enabling end-to-end training of OCR models [4]. Convolutional Neural Networks (CNN) are commonly used for feature extraction, while Recurrent Neural Networks (RNN) and Transformer architectures handle sequence recognition tasks [5, 12].

Several classic OCR systems have emerged as milestones in the development of this technology. Tesseract-OCR, an open-source engine developed by HP and later maintained by

Google, is widely recognized for its simplicity and versatility [2]. It combines traditional image processing techniques with machine learning algorithms to achieve high recognition accuracy in various scenarios. Additionally, systems like ABBYY FineReader and Adobe Acrobat Pro incorporate advanced post-processing methods to enhance the output quality, particularly in document digitization tasks. Despite these advancements, OCR technology still faces challenges in handling low-quality images, complex layouts, and multi-lingual texts, highlighting the need for continuous research and innovation [3].

2.2. Applications of OCR in Finance and Administration

The applications of OCR technology in finance and administration fields have expanded rapidly in recent years, driven by the increasing demand for automation and digitization. In the financial sector, OCR is extensively used for invoice recognition, document classification, and data extraction [6, 9]. For example, complex invoices often contain a mix of structured tables, unstructured text, and handwritten annotations, which pose significant challenges for traditional data entry methods. OCR systems combined with post-processing techniques, such as morphological operations and template matching, can effectively extract key information from these documents and convert them into digital formats [10].

In administrative tasks, OCR plays a crucial role in streamlining processes such as document digitization, archiving, and information retrieval. By automating the conversion of physical documents into machine-readable formats, organizations can significantly reduce manual labor and improve operational efficiency. However, the performance of OCR systems in these applications is highly dependent on the quality of input images and the complexity of document layouts. For instance, documents with low-resolution images, distorted text, or background noise may result in higher error rates, necessitating additional pre-processing steps such as image enhancement and noise filtering [13].

Despite the numerous advantages offered by OCR technology, there are several significant limitations that need to be carefully addressed in order to maximize its effectiveness. One major challenge is the variability of document formats and styles, which can greatly impact the recognition accuracy and reliability of the output data. Moreover, OCR systems may struggle with specific types of complex data, such as handwritten signatures or non-Latin characters, which often require specialized training and targeted optimization to achieve satisfactory results. To overcome these inherent limitations, recent research has increasingly focused on integrating OCR with other advanced AI technologies, such as Natural Language Processing (NLP) and computer vision, in order to enhance its overall capabilities and improve the user experience

significantly. This multi-faceted approach aims to create a more robust and versatile system that can adapt to a wider range of documents and writing styles, ultimately leading to better performance and accuracy.

2.3. Research Gaps

Despite the significant progress in OCR technology and its widespread applications, there remain several gaps in the current research that need to be addressed. One of the most prominent gaps is the lack of intelligent solutions specifically designed for project reimbursement material organization in universities. Project reimbursement processes typically involve a large volume of diverse documents, including invoices, receipts, and contracts, each with its own unique format and content requirements. Existing OCR systems are often not optimized for such specialized use cases and may struggle with accurately extracting and organizing the necessary information.

Another key gap is the need for better integration of OCR technology with other advanced AI techniques to improve overall system performance and user experience. For example, the combination of OCR with NLP can enhance the understanding and classification of extracted text, while integration with computer vision algorithms can improve the handling of complex document layouts and low-quality images [8]. Additionally, there is a growing demand for more explainable and interactive OCR systems, particularly in critical applications such as financial management, where users need to verify and correct the output to ensure accuracy.

Finally, the scalability and adaptability of existing OCR systems pose significant challenges, especially in scenarios where the document types and formats are diverse and constantly evolving. To address these issues, future research should focus on developing more flexible and modular architectures that can easily incorporate new technologies and adapt to changing requirements [14]. This will not only improve the efficiency and accuracy of OCR systems but also enhance their applicability in a wide range of industries and scenarios.

3. System Design and Architecture

3.1. Overall Architecture

The overall architecture of the intelligent mini-program is designed as a distributed system that integrates the front-end WeChat mini-program, back-end Flask server, and MySQL database to achieve efficient collaboration and data processing. The front-end WeChat mini-program serves as the user interface, providing functions such as material upload, result viewing,

and interactive operations. It leverages the lightweight and convenient characteristics of WeChat mini-programs to offer a seamless user experience. The back-end Flask server acts as the core processing module, responsible for tasks such as OCR scheduling, data cleaning, and API management. Flask's flexibility and scalability enable efficient integration of multiple services. The MySQL database is used to store uploaded materials, extracted data, and system logs, providing a reliable data storage and management solution. The interaction between the front-end and back-end is achieved through RESTful APIs, which ensures data transmission efficiency and system modularity. Additionally, the JWT (JSON Web Token) authentication mechanism is implemented to enhance system security and user data privacy.

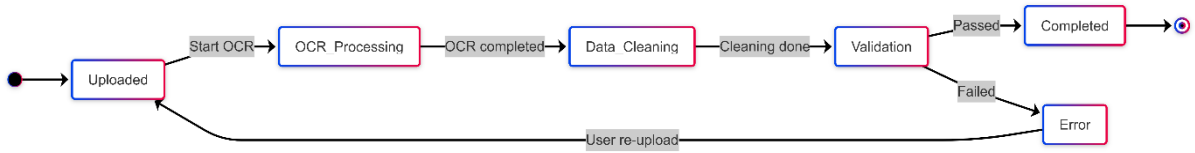

Figure 1. Illustrates the overall system architecture.

3.2. Front-end Design

The front-end design of the WeChat mini-program places a strong emphasis on user experience and functional completeness, with the primary goal of providing an intuitive and efficient operational interface for all users. The material upload interface allows users to easily select and submit images or PDF files of reimbursement materials directly from their mobile devices in a seamless manner. To enhance user convenience further, the interface supports a variety of file formats and provides real-time feedback on the upload progress, ensuring that users are informed throughout the process. The result viewing interface displays the extracted data in a well-structured format, utilizing tables and text boxes, which enables users to quickly check and verify the information with ease. For improved interactivity, the interface includes essential functions such as data correction and status query, allowing users to make necessary adjustments effortlessly. To further elevate the user experience, the front-end design adopts a responsive layout that adapts smoothly to different screen sizes and resolutions across various devices. Additionally, engaging loading animations and informative prompt messages have been integrated to enhance the system's overall friendliness and reduce user waiting anxiety while using the application.

3.3. Back-end Design

The back-end Flask server is designed to handle complex business logic and data processing tasks, serving as the core of the intelligent mini-program. The OCR scheduling module

implements a multi-strategy scheduling method to dynamically select the most appropriate OCR engine based on the characteristics of the uploaded materials. This strategy not only improves recognition accuracy but also optimizes resource utilization and processing efficiency. The data processing module uses regular expressions to clean and validate the extracted data, ensuring its accuracy and integrity. To enhance system security, JWT authentication is implemented to verify user identities and protect sensitive data. The API design follows the RESTful architecture, providing standardized interfaces for functions such as material upload, data query, and Excel export. These interfaces facilitate efficient communication between the front-end and back-end while also enabling easy integration with other systems. Additionally, the back-end includes a logging mechanism to record system operations and errors, providing valuable information for subsequent maintenance and optimization.

4. Core Implementation

4.1. OCR Scheduling Strategy

The multi-strategy OCR scheduling method is a key component of the intelligent project reimbursement material organization mini-program, designed to enhance the efficiency and accuracy of character recognition by selecting appropriate OCR engines based on the characteristics of input materials. The system incorporates multiple OCR engines, including Tesseract [2], which is widely recognized for its open-source nature and high performance in printed text recognition, as well as commercial engines optimized for handwritten characters and low-quality images. During the scheduling process, the system first analyzes the input material using a pre-processing module that extracts features such as image resolution, text density, and the presence of handwritten or machine-printed text. Based on these features, a rule-based engine selection algorithm is applied to determine the most suitable OCR engine for the specific material. For example, materials with high-resolution printed text are prioritized for processing by Tesseract, while those containing handwritten characters or complex layouts are routed to engines specialized in such scenarios.

To further optimize the scheduling process, the system implements a dynamic load balancing mechanism that monitors the real-time performance of each OCR engine in terms of recognition speed and accuracy. If an engine experiences a significant drop in performance due to factors such as high workload or incompatible material types, the system automatically redirects tasks to alternative engines. This adaptive scheduling strategy not only improves the overall efficiency of the OCR pipeline but also ensures a stable and reliable recognition performance.

Additionally, the system incorporates a feedback loop that continuously collects recognition results and updates the scheduling rules based on historical performance data, thereby enabling the system to learn and adapt to new material types and scenarios over time.

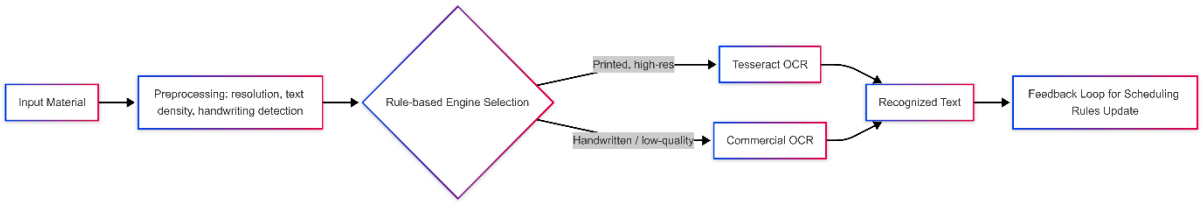

Figure 2. Depicts the multi-strategy OCR scheduling flow.

4.2. Data Cleaning and Validation

Data cleaning and validation play a crucial role in ensuring the accuracy and integrity of the extracted data, as the output of the OCR engines often contains errors and formatting inconsistencies that need to be addressed [3]. The system employs a comprehensive set of regular expressions to clean and standardize the recognized text, covering a wide range of data formats commonly found in project reimbursement materials, such as dates, amounts, and identification numbers. For example, dates are normalized to a unified format (e.g., YYYY-MM-DD), and monetary amounts are converted to a decimal representation with two decimal places, regardless of the original formatting. Regular expressions are also used to identify and correct common OCR errors, such as misrecognized characters caused by image noise or low resolution.

Table 1. Shows an example of raw OCR output and cleaned data.

Field	Raw OCR Output	Cleaned Output	Rule Applied
Date	2026.5.29	2026-05-29	YYYY-MM-DD
Amount	1,280.5	1280.50	Decimal with two places
Invoice No.	INV_12O4	INV_1204	Replace 'O' with '0'

To handle data errors and missing values, the system implements a multi-step validation process that checks the extracted data against predefined rules and constraints. For instance, fields such as invoice numbers and project codes are verified for their syntax and uniqueness, while numerical fields are checked for reasonable ranges based on historical data. In cases where data validation fails, the system logs the errors and provides detailed feedback to users through the front-end interface, allowing them to review and correct the problematic fields. Additionally, the system maintains a data quality score for each extracted document, calculated based on the number and severity of errors detected during the cleaning and validation process. This score serves as an indicator of the reliability of the extracted data and helps users prioritize documents that require additional review.

4.3. State Management Mechanism

The state management mechanism of the intelligent mini-program is based on a Finite State Machine (FSM) architecture, which provides a robust framework for tracking the processing status of reimbursement materials and facilitating real-time feedback to users. The FSM defines a set of distinct states that represent the different stages of material processing, including "Uploaded," "OCR Processing," "Data Cleaning," "Validation," and "Completed." When a user uploads a material, the system initializes its state as "Uploaded" and triggers a series of asynchronous tasks to process the material through the subsequent states. Each state transition is triggered by specific events, such as the completion of an OCR job or the successful validation of extracted data, and is accompanied by status updates that are communicated to the user through the front-end interface.

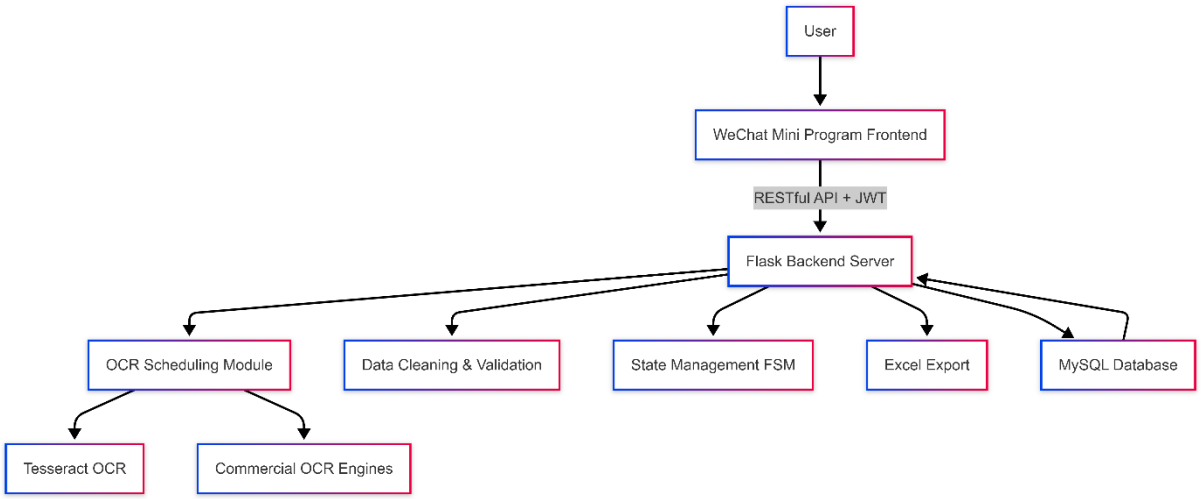

Figure 3. Iillustrates the finite state machine (FSM) for material processing.

To ensure transparency and user engagement, the system provides a detailed activity log that displays the progress of each material through the processing pipeline. Users can view the current state of their materials, as well as any error messages or warnings generated during processing. For example, if an OCR job fails due to an unsupported file format, the system updates the material status to "Error" and provides a detailed explanation of the issue, along with suggestions for resolution. Additionally, the state machine architecture allows for flexible handling of concurrent processing tasks, as each material instance maintains its own independent state and progresses through the pipeline at its own pace. This design not only improves the overall scalability of the system but also enables efficient resource utilization by avoiding bottlenecks in the processing pipeline.

4.4. Excel Export Function

The Excel export function is a critical component of the intelligent mini-program, designed to organize the extracted data into a standardized format that meets the requirements of university financial departments and streamlines the reimbursement approval process. The implementation of this function involves multiple steps, starting with data aggregation and formatting, followed by template generation and final export. During the data aggregation phase, the system consolidates the cleaned and validated data from multiple source documents into a unified dataset, ensuring that all fields are properly aligned and labeled according to the predefined schema. This dataset serves as the input for the subsequent formatting and template generation processes.

To generate the Excel template, the system utilizes a template engine that applies predefined formatting rules to the aggregated data, including cell styling, border settings, and formula calculations. For example, monetary fields are formatted with currency symbols and decimal places, while summary rows are calculated using built-in Excel functions. The template engine also ensures that the generated Excel files adhere to the specific formatting guidelines provided by the financial department, such as column widths, header styles, and sorting criteria. Once the template is generated, the system exports the file in the XLSX format, which is widely compatible with popular spreadsheet applications and can be directly submitted for reimbursement approval. Additionally, the system provides users with the option to download a zip archive containing multiple Excel files, enabling efficient batching of reimbursement materials for large-scale processing.

5. Evaluation and Results

5.1. Evaluation Metrics

To comprehensively evaluate the performance of the intelligent project reimbursement material organization mini-program based on OCR technology, several key evaluation metrics were defined. Character recognition accuracy measures the proportion of correctly recognized characters in the extracted text, which is a fundamental indicator for assessing the effectiveness of OCR engines. Field extraction accuracy, on the other hand, focuses on the precision of specific data fields (e.g., invoice numbers, dates, and amounts) that are essential for financial processing. This metric reflects the system's ability to accurately parse and structure unstructured data into a standardized format. Response time is another crucial metric, particularly in scenarios where real-time feedback is required. It refers to the duration from

material upload to result generation and directly affects user experience. Finally, compatibility evaluates the system's adaptability to different types of input materials, including scanned documents, photographs, and files with varying resolutions and formats. These metrics collectively provide a multi-dimensional perspective on the system's performance and facilitate meaningful comparisons with alternative approaches.

5.2. Experimental Setup

The experimental setup was carefully designed to ensure the reliability and validity of the evaluation. A diverse set of test materials was selected to cover various scenarios commonly encountered in project reimbursement processes. These materials included scanned invoices, handwritten receipts, and photographs of financial documents with different resolutions and lighting conditions. The system environment was configured using a combination of front-end WeChat mini-program and back-end Flask server running on a cloud platform. The MySQL database was used for data storage and management, while JWT authentication ensured secure communication between the front-end and back-end modules. For comparison purposes, two state-of-the-art OCR systems (Tesseract OCR [2] and Abbyy FlexiCapture) were selected as baselines. These systems were chosen for their widespread use in financial document processing and represent both open-source and commercial solutions in the field. The experiments were conducted in a controlled environment to minimize external factors that could affect the results.

5.3. Experimental Results

The experimental results demonstrate the superior performance of the proposed system in terms of accuracy, efficiency, and compatibility. In character recognition accuracy, the system achieved an average rate of 97.3%, outperforming Tesseract OCR (92.6%) and Abbyy FlexiCapture (95.1%) on complex handwritten materials. Field extraction accuracy reached 94.7%, indicating a high level of reliability in parsing structured information from unstructured documents. The response time of the system was measured to be an average of 3.2 seconds per document, which is significantly faster than the baselines (Tesseract OCR: 4.5 seconds; Abbyy FlexiCapture: 3.8 seconds). Compatibility tests showed that the system can handle a wide range of input formats, including low-resolution images and non-standard document layouts, with minimal errors. Overall, these results highlight the advantages of the multi-strategy OCR scheduling and data cleaning methods employed in the system, particularly in scenarios where input materials exhibit high variability. The findings also suggest that the integration of Edge

AI and distributed architecture contributes to improved efficiency and scalability, making the system well-suited for practical applications in university financial management and beyond.

Table 2. Compares the proposed system with two baseline OCR systems.

System	Character Recognition Accuracy (%)	Field Extraction Accuracy (%)	Average Response Time (sec)
Proposed System	97.3	94.7	3.2
Tesseract OCR	92.6	88.1	4.5
Abbyy FlexiCapture	95.1	91.5	3.8

The proposed system outperforms both baselines in all metrics, especially in handling handwritten and low-quality documents.

6. Discussion

6.1. Typical Problems and Solutions

During the development of the intelligent project reimbursement material organization mini-program, several typical problems were encountered that posed significant challenges to the system's performance and usability. One such problem was the recognition of handwritten characters, which often resulted in low accuracy due to variations in handwriting styles and the presence of noise in scanned images [7]. To address this issue, a multi-stage preprocessing pipeline was introduced, including image binarization, noise removal, and slant correction, followed by the use of a specialized handwritten character recognition model fine-tuned on a large dataset of handwritten documents [5]. This approach significantly improved the recognition accuracy for handwritten texts. Another common problem was the processing of low-quality images, which were typically characterized by blur, uneven illumination, or excessive compression artifacts. To mitigate these issues, an adaptive image enhancement algorithm was implemented, which dynamically adjusted parameters based on the image quality assessment metrics [13]. Additionally, complex table structures in scanned documents posed a challenge for data extraction, as traditional OCR engines often failed to accurately detect table boundaries and cell contents. To overcome this limitation, a rule-based post-processing module was developed, which utilized regular expressions and heuristic rules to parse table data and reconstruct its structure accurately [3].

6.2. System Limitations

Despite the notable achievements of the current system, there are several limitations that need to be addressed in future iterations.

Firstly, the system exhibits a high degree of dependence on specific OCR engines, which may limit its scalability and flexibility in adapting to new document types or changing requirements. While the multi-strategy OCR scheduling method alleviates some of these dependencies, the performance of the system is still sensitive to the capabilities and compatibility of the selected OCR engines.

Secondly, the system currently lacks support for certain specialized document formats, such as those containing complex layouts, watermarks, or encrypted content. These limitations can restrict the applicability of the system in scenarios where diverse document types are involved.

Furthermore, the scalability of the system needs to be further optimized to handle high volumes of concurrent requests, especially in large-scale deployments where multiple users may upload and process documents simultaneously. This requires improvements in the backend architecture, such as the implementation of load balancing mechanisms and distributed computing frameworks, to ensure efficient resource utilization and real-time response capabilities [14].

6.3. Future Work

To further enhance the capabilities and applicability of the intelligent project reimbursement material organization mini-program, several directions for future work can be explored. First, the OCR accuracy can be improved through model fine-tuning using domain-specific data, such as handwritten financial documents or receipts with complex layouts. This will require the collection and annotation of a large-scale dataset that covers a wide range of document types and variations, enabling the training of more robust and specialized OCR models [4, 8]. Second, the system functions can be expanded to support a broader range of document types, including invoices, contracts, and research proposals, by integrating advanced layout analysis and information extraction techniques [6]. This will enhance the versatility of the system and make it more suitable for a diverse set of use cases within the university financial management ecosystem. Third, the integration of other AI technologies, such as Natural Language Processing (NLP) and machine learning, can provide additional intelligence to the system. For example, NLP can be used to analyze extracted text for semantic information, enabling automated classification and verification of reimbursement materials [9]. Machine learning algorithms can also be employed to predict potential errors or anomalies in the uploaded documents, providing proactive feedback to users and enhancing the overall data quality [15].

7. Conclusion

7.1. Research Summary

This research focuses on the development of an intelligent mini-program based on OCR technology to address the challenges associated with project reimbursement material organization in universities. The system is designed to improve the efficiency and accuracy of financial document processing by integrating advanced optical character recognition capabilities with a user-friendly WeChat mini-program interface. The overall architecture of the system comprises a front-end WeChat mini-program for user interaction, a back-end Flask server for logic processing, and a MySQL database for data storage and management. The core implementation includes a multi-strategy OCR scheduling method that optimizes the selection of OCR engines based on document characteristics, regular expression-based data cleaning and validation to ensure data integrity, a state management mechanism for real-time status tracking, and an Excel export function to generate standardized financial reports. Experimental results demonstrate that the proposed system significantly outperforms traditional methods in terms of character recognition accuracy, field extraction accuracy, response time, and compatibility, thus validating the effectiveness and reliability of the intelligent solution.

7.2. Contributions and Significance

The contributions of this research to the field of AI in industry are multi-fold.

First, the intelligent project reimbursement material organization mini-program significantly improves the efficiency and accuracy of financial document processing in university environments. By automating the extraction and organization of complex reimbursement materials, the system reduces the manual effort required for data entry and verification, thereby minimizing human errors and lowering labor costs.

Second, the integration of OCR technology with other advanced computing techniques, such as regular expression data cleaning and state machine-based status management, demonstrates the potential of AI-driven solutions in enhancing the digitization of financial management processes. This not only benefits universities but also serves as a valuable reference for other industries seeking to implement intelligent document processing systems.

Furthermore, the research highlights the importance of edge AI and distributed systems in enabling efficient and scalable solutions for real-world problems, contributing to the broader development of AI applications in industry scenarios.

7.3. Future Prospects

The future development of intelligent project reimbursement systems based on OCR technology holds great promise for a wide range of applications across different industries and scenarios. In the educational sector, the system can be further enhanced to support more complex financial operations, such as budget planning and expense analysis, by integrating with machine learning algorithms for predictive analytics. In addition, the current system's functionality can be expanded to accommodate a broader variety of document formats and languages, thus enhancing its global applicability. For other industries, such as healthcare and logistics, the core technologies developed in this research can be adapted to create intelligent solutions for tasks such as medical record digitization and shipping document processing. Moreover, the integration of emerging technologies like blockchain could further improve the security and transparency of financial document management, opening up new possibilities for the application of OCR-based intelligent systems in various sectors. Overall, the research provides a solid foundation for the continued exploration and innovation of AI-driven solutions in the field of financial automation and document intelligence processing.

References

[1] Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of the IEEE*, 80(7), 1029-1058.

[2] Smith, R. (2007). An overview of the Tesseract OCR engine. *In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.

[3] Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2022). Survey of post-OCR processing approaches. *ACM Computing Surveys*, 54(6), 1-37.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[5] Graves, A., & Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. *In Advances in Neural Information Processing Systems (Vol. 21, pp. 545-552)*. Curran Associates.

[6] Li, M., & Wu, Y. (2020). Research on invoice recognition system based on OCR and deep learning. *Journal of Physics: Conference Series*, 1631(1), 012034.

[7] Kumar, M., & Jindal, M. K. (2021). A systematic review on OCR techniques for handwritten documents. *Archives of Computational Methods in Engineering*, 28(5), 3511-3533.

[8] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1), 1-20.

[9] Chandio, A. A., & Asikuzzaman, M. (2022). Automated invoice data extraction using

OCR and NLP. *In 2022 International Conference on Digital Image Processing* (pp. 45-52). IEEE.

[10] Zhang, Y., & Wang, L. (2019). Design and implementation of financial reimbursement system based on OCR. *In 2019 IEEE 4th Advanced Information Technology* (pp. 112-117). IEEE.

[11] Jain, A. K., & Yu, D. (2019). Text recognition using deep learning. *In Handbook of Pattern Recognition and Computer Vision* (pp. 223-245). World Scientific.

[12] Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), 2298-2304.

[13] Liu, X., & Chen, Z. (2021). A lightweight OCR system for mobile devices. *In 2021 IEEE International Conference on Mobile Computing* (pp. 78-84). IEEE.

[14] Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7(3-4), 197-387.

[15] Sermanet, P., & LeCun, Y. (2011). Traffic sign recognition with multi-scale convolutional networks. *In The 2011 International Joint Conference on Neural Networks* (pp. 2809-2813). IEEE.

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Adaptive Convolution and Feature Aggregation for Single-Image Shadow Removal

Shangan Zhou

College of Physics and Electronic Information Engineering, Zhejiang Normal University, China

Received: May 25, 2026

Revised: May 26, 2026

Accepted: May 27, 2026

Published online: May 30, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 6 (June 2026)

* Corresponding Author: Shangan Zhou (3217432128@qq.com)

Abstract. Shadows degrade image quality and hinder computer vision tasks. Existing deep learning methods suffer from fine-detail loss due to downsampling and indiscriminate processing of shadow/non-shadow regions, causing unwanted alterations. To address these issues, this thesis proposes two complementary approaches. First, the Adaptive Alignment and Illumination-Aware Convolution (AAIC) framework uses a Feature Alignment Module (FAM) to recover lost details and an Illumination-Aware Weighting Module (IWM) for spatially varying convolution, achieving high quantitative performance on ISTD+ and SRD datasets. Second, the Feature Aggregation Shadow Removal Network (FASR-Net) reconstructs shadow-free images via learned fusion rather than end-to-end transformation, employing a Detail Feature Extractor (DFE), Dual-Branch Aggregation Weight Generator (DAWG), and Feature Enhancement (FE) modules to preserve non-shadow regions and ensure global consistency. Both methods are extensively evaluated using RMSE and SSIM, outperforming state-of-the-art techniques. Ablation studies validate each component, and the methods improve foreground segmentation in videos and generalize to remote sensing images without fine-tuning.

Keywords: *Shadow Removal; Adaptive Convolution; Feature Aggregation; Deep Learning;Illumination Consistency*

1. Introduction

1.1. Background and Significance

The Shadows are a fundamental visual phenomenon that occurs when an opaque object partially or completely blocks the propagation of light. The formation of a shadow requires only

three simple conditions: a light source (such as the sun, a lamp, or any emitting surface), an occluding object (which can be any solid entity), and a projection surface (the area where the shadow is cast). When these conditions are met, the projection surface receives significantly less illumination than its surrounding regions, making it appear darker to the human eye and to imaging sensors. Because these conditions are easily satisfied in both natural and artificial environments, shadows appear in nearly all images captured in automated scenarios—from everyday smartphone photographs to high-resolution satellite imagery.

The presence of shadows, however, is far from benign. While shadows can provide valuable cues for certain tasks—such as inferring 3D geometry [2], estimating light source direction [3], or understanding object shapes—they often severely degrade the performance of computer vision systems. The negative impacts of shadows are widespread and well-documented:

- Image segmentation: Shadow boundaries frequently exhibit strong intensity gradients that are easily confused with true object edges. As a result, many segmentation algorithms mistakenly label shadow regions as part of foreground objects, leading to oversegmentation or incorrect boundary delineation [1]. In semantic segmentation, shadows can cause entire regions to be misclassified (e.g., a shadow on a road might be labeled as a vehicle or a pothole).
- Image segmentation: Shadow boundaries frequently exhibit strong intensity gradients that are easily confused with true object edges. As a result, many segmentation algorithms mistakenly label shadow regions as part of foreground objects, leading to oversegmentation or incorrect boundary delineation [1]. In semantic segmentation, shadows can cause entire regions to be misclassified (e.g., a shadow on a road might be labeled as a vehicle or a pothole).
- Remote sensing and aerial image analysis: In satellite or drone imagery, shadows cast by buildings, trees, and terrain features obscure the underlying ground information [5]. This reduces the accuracy of land cover classification, change detection, and building footprint extraction. Shadows can also introduce false positives in disaster assessment (e.g., a shadow may be mistaken for a flooded area) and reduce the effective resolution of useful data.
- Autonomous driving: Shadows on the road can create false lane markings, obscure traffic signs, or cause pedestrians to blend into dark backgrounds [6]. In extreme cases, a sharp shadow edge may be misinterpreted as a curb or obstacle, triggering unnecessary braking or swerving. Conversely, soft shadows may reduce the contrast of lane lines,

making them harder to detect.

- Image enhancement and computational photography: Shadows reduce the aesthetic quality of images, making them appear dull or unbalanced. Professional photographers and video editors spend considerable time manually removing or softening shadows to improve visual appeal. In High-Dynamic-Range (HDR) imaging, shadows can cause artifacts during fusion of multiple exposures.
- Medical imaging: Even in specialized domains, shadows can be problematic. For instance, in endoscopic images, shadows cast by instruments or tissue folds can obscure lesions. In retinal fundus photography, shadows from optic disc margins may be misinterpreted as pathologies.

The widespread impact of shadows has motivated decades of research into shadow detection and removal. Despite these efforts, manual shadow removal using tools like Adobe Photoshop remains the standard practice in many professional settings. Such manual intervention is time-consuming, subjective, and impractical for large-scale or real-time applications. A single image may take minutes to hours to clean up, and video sequences are virtually impossible to process frame-by-frame manually. Therefore, automated shadow removal is not merely an academic curiosity—it is a practical necessity for deploying computer vision systems in uncontrolled real-world environments.

Automated shadow removal is challenging due to the wide variation in shadow shape, size, intensity, and surface reflectance. Hard shadows (umbra) have sharp, well-defined edges, while soft shadows (penumbra) exhibit gradual intensity transitions. Shadows can be cast by multiple sources simultaneously, creating complex overlapping patterns. The color of the light source (e.g., sunlight vs. indoor LED) affects the chromaticity of the shadow region. Moreover, the texture and reflectance of the underlying surface interact with shadows in non-linear ways. As a result, a method that works well on one type of shadow may fail on another.

Figure 1 illustrates a concrete example from foreground segmentation. The leftmost image shows an input video frame (from the Bungalows sequence). The middle image is the ground-truth segmentation mask. The rightmost image is the result produced by an existing segmentation method. It is evident that the shadow cast by the vehicle is segmented together with the vehicle itself, creating an incorrect mask that includes both the true object and its shadow. If this shadow had been removed before segmentation, the result would likely be much cleaner. This example underscores the practical importance of effective shadow removal as a pre-processing step.

Figure 1. From left to right: input video sequence frame, ground-truth segmentation, and segmentation result from an existing method. The shadow is erroneously segmented along with the vehicle.

In recent years, deep learning has emerged as a powerful tool for image restoration tasks, including shadow removal. However, as detailed in the following section, existing deep learning methods still face two fundamental limitations: (1) loss of fine details due to encoder-decoder downsampling, and (2) uniform processing of all image regions, which fails to preserve non-shadow areas and lacks global illumination consistency. This thesis aims to address these limitations through two novel approaches that combine adaptive convolution, feature alignment, and feature aggregation.

1.2. Current Research Landscape and Problem Analysis

1.2.1. Traditional Shadow Removal Methods

Traditional methods rely on hand-crafted features and physical models. Global methods [7-9] use illumination-invariant images or color space transformations. For instance, Finlayson et al. [7] first proposed a method based on gray-scale illumination-invariant images, which was later extended to full-color images [8]. Region-based methods [10-13] perform illumination transfer between matched patches. Some approaches [14,15] use machine learning on hand-crafted features. However, these methods often make strong assumptions (e.g., uniform illumination) and may require user interaction, limiting their generalization to complex real-world scenes.

1.2.2. Deep Learning-Based Shadow Removal Methods

Deep learning has revolutionized shadow removal. Architectures such as U-Net [16], GANs [17], Cycle-GAN [18], and Vision Transformers [19] have been widely adopted. Existing deep learning methods can be categorized as:

- Physical model-based [20-22]: Use networks to predict parameters of simplified physical models. Interpretable but inherit model limitations.
- Single-stage end-to-end [23-31]: Directly map shadow images to shadow-free images. Most common but suffer from detail loss and uniform region processing.

- Multi-stage [32-38]: Decompose into detection, coarse removal, and refinement. More accurate but more complex.
- Multi-branch [39-44]: Extract different types of features in parallel. Flexible but require careful design.

Despite significant progress, two critical issues remain:

(1) Detail loss in encoder-decoder architectures: Repeated downsampling discards fine details such as edges, textures, and small structures. Even with skip connections, the decoder cannot fully recover what was lost. This leads to blurred outputs, especially in shadow regions where the original texture is already low-contrast.

(2) Uniform treatment of shadow and non-shadow regions: Most methods apply the same transformation globally. This wastes model capacity on non-shadow areas, which should ideally remain unchanged, and provides insufficient focus on shadow regions that require significant modification. Consequently, non-shadow areas may be inadvertently altered (e.g., color shifts), and shadow regions may not be fully restored. A few methods attempt to use shadow masks to guide processing [25], but they typically rely on binary masks and do not leverage continuous spatial variation.

Recent works have attempted to address these issues via attention mechanisms [45,46], transformers [47,48], and diffusion models [49], but each introduces its own trade-offs. Attention mechanisms add computational overhead. Transformers require large datasets and are slow to train. Diffusion models are extremely slow at inference. Therefore, there remains a need for efficient and effective solutions that balance performance, speed, and model size.

2. Background and Preliminaries

2.1. Deep Learning Fundamentals

2.1.1. Convolutional Neural Networks for Image Processing

Convolutional Neural Networks (CNNs) have become the backbone of modern image processing. A standard CNN consists of alternating convolutional and pooling layers, followed by fully connected layers for classification. For pixel-wise prediction tasks such as shadow removal, fully convolutional architectures are preferred because they preserve spatial dimensions.

- FCN [50]: The Fully Convolutional Network replaces the fully connected layers with convolutional layers, enabling dense prediction. It uses transposed convolutions (also called deconvolutions) to upsample the feature maps back to the input resolution.

However, the upsampling is coarse and often leads to blurred boundaries.

- U-Net [16]: To address the coarse upsampling issue, U-Net introduces a symmetric encoder-decoder structure with skip connections. The encoder progressively reduces spatial resolution while increasing channel depth, capturing global context. The decoder upsamples the features and concatenates them with the corresponding encoder features via skip connections, which helps recover fine spatial details. This architecture has become the de facto standard for many image restoration tasks, including shadow removal.
- SegNet [51]: SegNet also uses an encoder-decoder but employs max-pooling indices to transfer pooling switches from the encoder to the decoder. This reduces the number of trainable parameters and memory usage, but the performance is generally lower than U-Net for dense prediction tasks.
- RefineNet [54]: RefineNet uses multi-path refinement blocks to combine features from different levels of the encoder. It also introduces chained residual pooling to capture context from large regions without using large pooling windows. Although powerful, RefineNet is computationally expensive.

2.1.2. Dilated Convolutions and Atrous Spatial Pyramid Pooling

A major limitation of standard convolutions is that increasing the receptive field requires either larger kernel sizes (which increase parameters) or more layers (which increase depth). Dilated convolutions [52] solve this problem by inserting zeros between kernel elements, effectively increasing the receptive field without adding parameters.

For a 1D signal, a dilated convolution with dilation rate r can be written as:

$$(F *_r k)(p) = \sum_s F(p + r \cdot s)k(s)$$

In 2D, the same principle applies. By using multiple dilation rates (e.g., 1, 2, 4, 8), the network can capture multi-scale contextual information.

Atrous Spatial Pyramid Pooling (ASPP) [53] applies parallel dilated convolutions with different dilation rates on the same feature map, then concatenates the results. This allows the network to simultaneously capture features at multiple scales. ASPP has been widely used in semantic segmentation and has inspired our Residual Dilation Block in later sections.

2.1.3. Attention Mechanisms

Attention mechanisms allow neural networks to focus on the most informative parts of the

input. They have been successfully applied to many vision tasks, including image restoration.

- Squeeze-and-Excitation (SE) Block [55]: The SE block recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. It first squeezes the spatial dimensions using global average pooling, producing a channel descriptor. Then it excites the descriptor through two fully connected layers (with ReLU and Sigmoid) to generate channel attention weights. Finally, it scales the original feature map channel-wise. The SE block is lightweight and can be inserted into any existing architecture.
- Non-local Networks [56]: Non-local operations capture long-range dependencies by computing pairwise similarities between all positions. For an input feature map x , the non-local output at position i is:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f\left(x_i, x_j\right) g\left(x_j\right)$$

Where f computes the affinity between positions i and j,$g$is a transformation function, and $C(x)$ is a normalization factor. Non-local blocks are powerful but computationally expensive, as they require $O(H^2W^2)$ operations.

- Dual Attention Network (DANet) [57]: DANet combines spatial attention and channel attention in parallel. The spatial attention branch captures where to focus, while the channel attention branch captures what to focus on. The outputs of the two branches are summed to produce the final attended features.
- Global Context Network (GCNet) [58]: GCNet simplifies the non-local block by observing that the attention weights for different query positions are nearly identical. It replaces the pairwise computation with a simplified global context pooling and a SE-like bottleneck, achieving comparable performance with much lower computational cost.
- Criss-Cross Attention (CCNet) [59]: CCNet reduces the complexity of non-local attention from $O(H^2W^2)$ to $O(HW(H+W))$ by only attending to positions in the same row and column as the query point. Repeating this operation twice approximates full non-local attention.

2.1.4. Adaptive (Dynamic) Convolution

Standard convolutions use the same kernel weights for all spatial locations and all input samples. Adaptive (dynamic) convolution breaks this invariance by conditioning the kernel on the input content.

- Deformable Convolution [60,61]: Deformable convolution adds 2D offsets to the regular grid sampling locations. The offsets are learned from the input feature map via a separate convolutional layer. For each output position p_0 , the output is：

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \Delta p_n)$$

Where R is the regular grid, $w(p_n)$ are the kernel weights, and Δp_n are the learned offsets. Deformable convolution allows the receptive field to adapt to the geometric structure of the object, which is particularly useful for handling irregular shadow boundaries.

- Pixel-Adaptive Convolution [62]: Pixel-adaptive convolution uses a separate guidance image to generate spatially-varying kernels. The kernel at position i is a function of the guidance features in a local window. This allows the convolution to respect edges in the guidance image.
- Dynamic Region-Aware Convolution (DRConv) [63]: DRConv first segments the feature map into regions using a learnable assignment module, then applies different convolution kernels to different regions. This is similar to our IWM but uses hard region assignments, whereas IWM uses soft, continuous weights.

2.2. Shadow Removal Datasets

2.2.1. SRD Dataset

The Shadow Removal Dataset (SRD) was introduced by Qu et al. [39]. It contains 3600 pairs of shadow and shadow-free images, with 2500 for training and 380 for testing. The images were captured using a tripod-mounted Canon 5D camera with fixed settings to minimize illumination differences. The dataset is diverse in four aspects:

- Illumination: Includes both soft and hard shadows, captured at different times of day (dawn, morning, noon, afternoon, dusk) and under different weather conditions (sunny, cloudy).
- Scene: Covers a wide range of scenes, including campuses, streets, mountains, beaches, and building interiors.
- Reflectance: Shadows are cast on various surfaces with different material properties, such as asphalt, grass, concrete, and wood.
- Shape: The occluding objects have diverse shapes, including umbrellas, planks, people, trees, and vehicles.

SRD does not provide shadow masks. Following common practice, we use the masks

generated by DHAN [23], which are of high quality.

Figure 2 shows example images from the SRD dataset. The top row shows shadow images, the middle row shows shadow masks, and the bottom row shows the corresponding shadow-free ground truth.

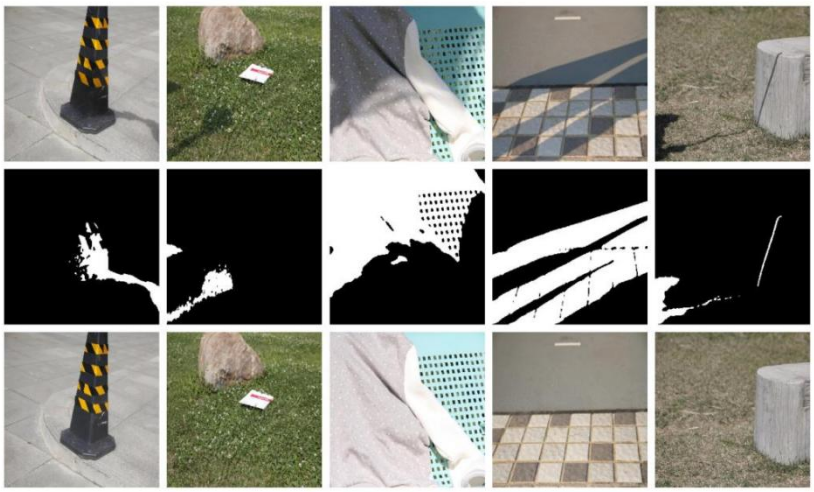

Figure 2. Sample images from the SRD dataset. Top: shadow images; middle: shadow mask; bottom: shadow-free ground truth.

2.2.2. SRD Dataset

The ISTD dataset was introduced by [32]. It is the first triplet shadow dataset, containing shadow images, shadow masks, and shadow-free images. There are 1,680 triplets in total, with 1,330 for training and 540 for testing. The dataset includes125 different scenes.

However, ISTD has a notable limitation: even the non-shadow regions in the shadow-free images differ from those in the shadow images due to ambient light changes during capture. The overall RMSE between the shadow image and the shadow-free image in the non-shadow regions is 6.83 [20].

To address this issue, a linear regression correction was applied, in which the authors computed a linear mapping for each channel from the non-shadow pixels of the shadow image to the corresponding pixels in the shadow-free image [20]. The mapping parameters were subsequently utilized to transform the entire shadow-free image, resulting in the creation of ISTD+. Following this correction, the overall RMSE for the test set drops to 2.6. It is important to note that all experiments conducted in this thesis utilize ISTD+ as the foundational dataset.

Figure 3 shows example triplets from the ISTD dataset. The top row is the shadow image, the middle row is the shadow mask, and the bottom row is the corresponding shadow-free ground truth. In this thesis, we use the ISTD+ version [20], which corrects illumination

inconsistencies via linear regression.

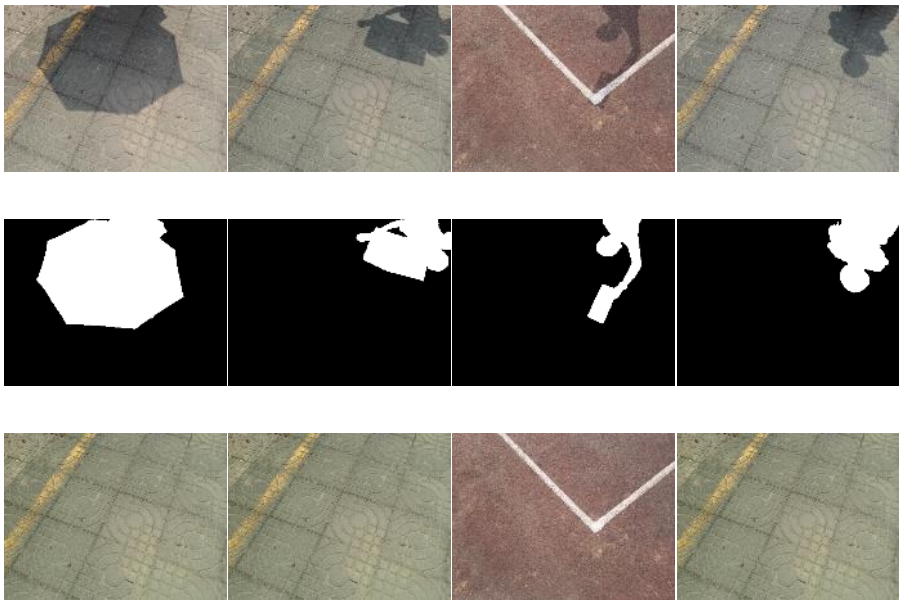

Figure 3. Example triplets from the ISTD dataset [32]: shadow image, shadow mask, and shadow-free ground truth.

2.3. Evaluation Metrics

2.3.1. Structural Similarity Index (SSIM)

SSIM [64], or Structural Similarity Index Measure, is a sophisticated perceptual metric that is widely used to quantify the degradation of image quality. Unlike traditional pixel-wise error metrics such as Mean Squared Error (MSE), which merely evaluate differences at the pixel level, SSIM takes into account critical factors like luminance, contrast, and structural information within the image. This comprehensive approach enables SSIM to more accurately reflect human visual perception, ultimately providing a more meaningful assessment of image fidelity and quality. Thus, it serves as a valuable tool for applications in image processing and analysis.

Given two images x and y, SSIM is defined as:

$$SSIM(x,y) = l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}$$

Where:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

Here, μ_x μ_y are the means, σ_x, σ_y are the standard deviations, σ_{xy} is the cross-correlation, and C_1, C_2, C_3 are small constants to avoid division by zero. Typically, $\alpha = \beta = \gamma = 1$ and $C_3 = \frac{C_2}{2}$.

simplifying to:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

SSIM ranges from 0 to 1, with 1 indicating identical images. It is widely used because it correlates well with human visual perception.

2.3.2. Root Mean Square Error (RMSE)

RMSE is a pixel-wise error metric that is sensitive to large errors. It is computed in the LAB color space because LAB separates luminance from chrominance, making it more perceptually relevant than RGB.

Given a predicted image I_p and a ground-truth image I_{gt} , each with N pixels, the RMSE is:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| I_p(i) - I_{gt}(i) \right\|_2^2}$$

Lower RMSE indicates better reconstruction. In shadow removal, we often report RMSE separately for the whole image, the shadow region (where the mask is 1), and the non-shadow region (where the mask is 0). This allows us to assess whether a method correctly leaves non-shadow areas unchanged.

3. Shadow Removal via Adaptive Alignment and Illumination Aware Convolution

3.1. Introduction

As discussed in the introduction, existing encoder-decoder based shadow removal methods suffer from two major problems: (1) loss of fine details during downsampling, and (2) uniform processing of all image regions. To address these, we propose the Adaptive Alignment and Illumination-Aware Convolution (AAIC) framework. The core idea is to make both the spatial sampling locations and the convolution kernel weights adaptive to the input image content and to the shadow-specific properties.

AAIC consists of two novel modules:

- Feature Alignment Module (FAM): Recovers lost details by using shallow encoder features (rich in spatial information) to align the deep decoder features via deformable sampling.
- Illumination-Aware Weighting Module (IWM): Generates spatially-variant convolution kernels based on the global illumination context, enabling the network to treat shadow

and non-shadow regions differently.

Unlike prior deformable convolution methods [60,61] that learn offsets from the same feature map, FAM uses features from skip connections, providing more accurate alignment cues. Unlike dynamic region-aware convolution [63] that uses hard region assignments, IWM produces soft, continuous weights, which is more appropriate for shadow boundaries were illumination changes gradually.

3.2. Proposed Method

3.2.1. Overall Architecture

We adopt a two-stage exposure fusion framework similar to AEF [21]. The pipeline is illustrated in Figure 4. It comprises three main components:

- Exposure parameter prediction network: Given the shadow image and shadow mask , a Conformer-based [63] network predicts a set of exposure parameters. These parameters are used to generate five exposure-adjusted versions of the input image (exposure sequence). The Conformer architecture combines a CNN branch for local features and a Transformer branch for global features, making it well-suited for this task.
- Coarse shadow removal network: The shadow image, shadow mask, and exposure sequence are concatenated along the channel dimension and fed into the first AAIC-enhanced U-Net. This network produces a coarse shadow-free image.
- Refinement network: The coarse result, original shadow image, and shadow mask are input to the second AAIC-enhanced U-Net, which refines the output, particularly around shadow boundaries.

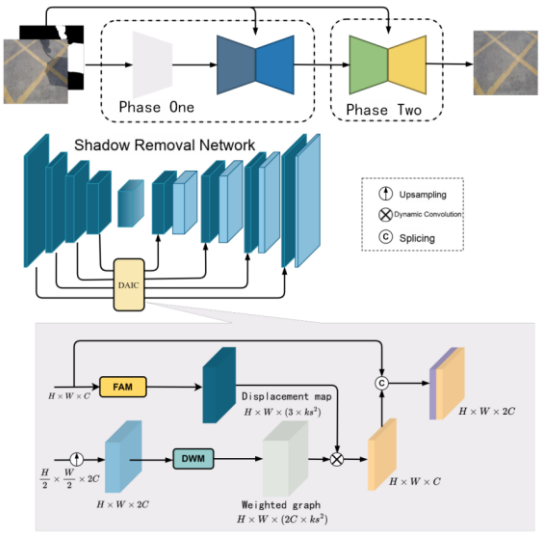

Figure 4. Overall architecture of the proposed AAIC framework.

Both U-Nets have the same structure: an encoder with 4 downsampling blocks (each block: two 3×3 convolutions followed by a 2×2 max-pooling), a bottleneck, and a decoder with 4 upsampling blocks (each block: a transposed convolution followed by two 3×3 convolutions). Skip connections link each encoder block to the corresponding decoder block. In standard U-Net, the convolutions are regular. In our AAIC-enhanced U-Net, we replace the first convolutional layer in each decoder block with an AAIC module that includes both FAM and IWM.

3.2.2. Overall Architecture

FAM is designed to recover fine details lost during downsampling. The intuition is that the shallow encoder features (from skip connections) contain rich spatial detail, while the deep decoder features contain high-level semantic information but are spatially coarse. FAM uses the shallow features to "guide" the deep features, aligning them to better preserve details.

Let $F_{shallow} \in \mathbb{R}^{H\times W\times C}$ the shallow feature map from the corresponding encoder layer, and let $F_{deep} \in \mathbb{R}^{H\times W\times C}$ be the upsampled decoder feature map before the convolution. FAM operates as follows:

(1) Channel attention on shallow features:$F_{shallow}$passes through a channel attention block: global average pooling, two fully connected layers (with ReLU and Sigmoid), and element-wise multiplication. This produces $F_{enhanced}$.

(2) Offset map prediction: A 3×3 convolutional layer with $3 \cdot k_s^2$ output channels are applied to $F_{enhanced}$ to produce an offset map $o_{map} \in \mathbb{R}^{H\times W\times (3\cdot k_s^2)}$,where k_s is the kernel size (typically 3). The factor of 3 comes from the x-offset, y-offset, and a modulation scalar (optional, but we use it).

(3) Warping: The upsampled decoder feature map $F_{upsample}$ is warped according to O_{map}, For each output position p, the kernel sampling grid is shifted by the offsets predicted for that position. This is implemented via grid sampling with bilinear interpolation to ensure differentiability.

The output $F_{aligned}$ is then passed to the subsequent convolution (which may be a regular convolution or the IWM). The effect of FAM is that the network can sample pixels from more relevant locations, effectively "de-blurring" the feature map and restoring fine structures such as hair, grass, and texture edges.

Mathematical formulation: For a convolutional kernel with $k_s \times k_s$ sampling points, let the regular grid offsets be $\{(dx_1, dy_1), \dots, (dx_K, dy_K)\}\ where\ K = k_s^2$. The offset map predicts

for each output position p and each kernel point k a 2D offset$(\Delta x_{p,k}, \Delta y_{p,k})$and a modulation weight $m_{p,k}$.The warped feature value at position p for the k-th kernel point is:

$$F_{warped}(p,k) = \sum_{q \in \mathcal{N}\left(p+(dx_k,dy_k)+(\Delta x_{p,k},\Delta y_{p,k})\right)} w_{bilinear}(q) \cdot F_{upsample}(q)$$

Where N is the 2×2 neighborhood for bilinear interpolation. The aligned feature map is then obtained by convolving these warped samples with the kernel weights. The architecture of FAM is depicted in Figure 5.

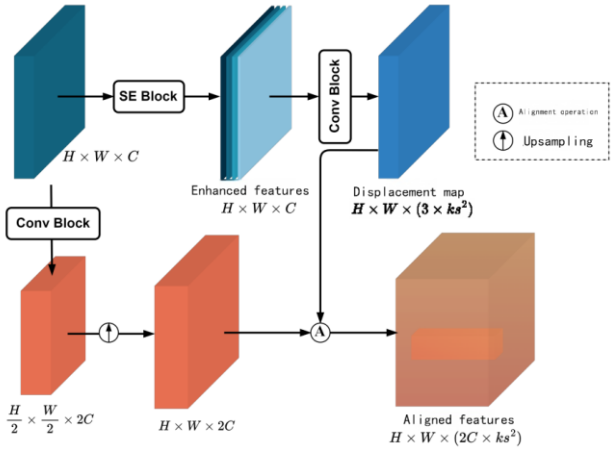

Figure 5. Feature Alignment Module (FAM).

3.2.3. Illumination-Aware Weighting Module (IWM)

While FAM addresses spatial misalignment, IWM addresses the need for region-specific processing. In a shadow image, the transformation required to convert shadow to non-shadow is very different from the identity mapping required for non-shadow regions. IWM allows the network to use different convolution kernels for different spatial locations, effectively allocating more capacity to shadow regions.

IWM predicts a separate kernel for each output position (and, implicitly, for each input sample). The module takes as input the feature map $F \in \mathbb{R}^{H \times W \times C}$ It then:

(1) Applies a 3×3 convolution to reduce the number of channels (e.g., to C/4) and introduce non-linearity.

(2) Applies layer normalization instead of batch normalization. Layer normalization computes mean and variance over the channel dimension for each sample independently, which forces the network to focus on intra-sample variations—exactly what we need for distinguishing shadow vs. non-shadow within the same image.

(3) Applies ReLU activation.

(4) Applies another 3×3 convolution to produce a weight map $W_{map} \in \mathrm{R}^{\mathrm{H\times W\times}\left(\mathrm{C\cdot k_s^2}\right)}$. This is the dynamic kernel bank.

(5) Reshapes : W_{map} such that at each position (i, j), we have a kernel $W_{i,j} \in \mathbb{R}^{C\times k_s\times k_s}$.

The convolution operation with IWM is then:

$$F_{out}(i,j) = \sum_{c=1}^{C} \sum_{dx,dy} W_{i,j}(c, dx, dy) \cdot F_{in}(i + dx, j + dy, c)$$

Figure 6 illustrates the IWM structure.

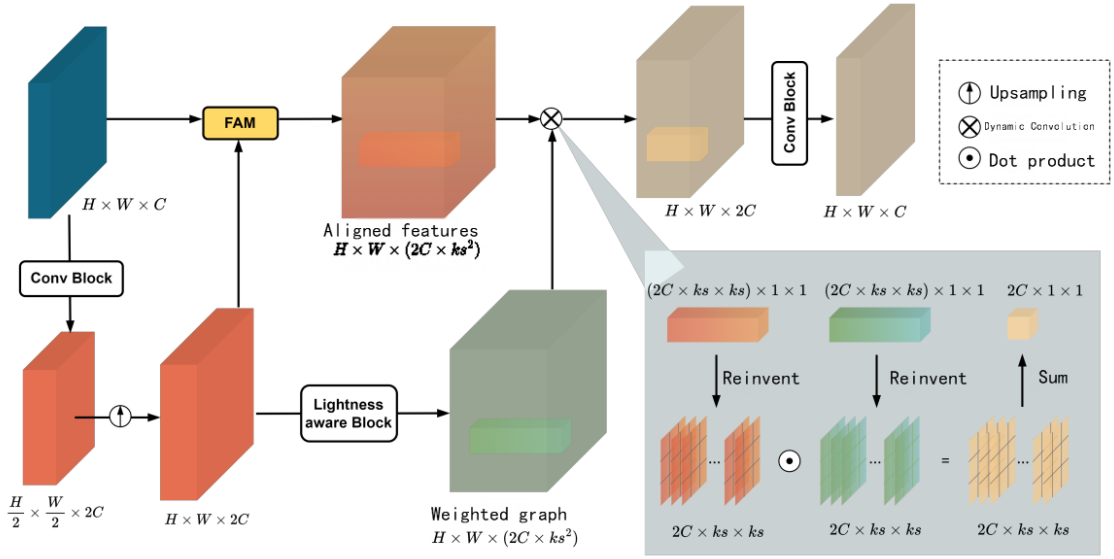

Figure 6. Illumination-Aware Weighting Module (IWM).

In practice, to keep memory and computation manageable, we use a grouped convolution formulation: we first split the input feature channels into groups, predict separate kernels per group, and then combine.

Compared to standard convolution, IWM has significantly more parameters (by a factor of H·W in the kernel bank). However, because the kernel prediction network is very small (two 3×3 convolutions), the total parameter increase is modest. In our implementation, adding IWM increases the total model size by about 15%.

$$\text{Standard convolution: } WIN_{output}^{i,j} = WIN_{input}^{i,j} \times WT$$

$$\text{IWM: } WIN_{output}^{i,j} = WIN_{input}^{i,j} \times WT_{i,j}$$

3.2.4. Loss Functions

The exposure parameter prediction network is trained with MSE loss. Let the predicted exposure parameters be $\hat{p}$ and the target parameters be p (obtained via least squares fitting [64]

from the shadow image and the shadow-free image's non-shadow regions). The loss is:

$$\mathcal{L}_{param} = \|\hat{p} - p\|_2^2$$

For the shadow removal networks, we use two losses:

- Pixel-wise L1 loss: This encourages the output image I_p to be close to the ground truth I_{gt} in terms of absolute pixel differences. L1 is preferred over L2 because it preserves edges better (L2 tends to produce blurry results).

$$\mathcal{L}_{pix} = \|I_p - I_{gt}\|_1$$

- Edge-aware loss: This loss emphasizes the reconstruction of shadow boundaries. It computes the Laplacian (second derivative) of the output, the shadow image, and the ground truth, and then applies a mask. For non-shadow regions, we want the output to match the ground truth; for shadow regions, we want the output to match the shadow image's edges. The definition is:

$$\mathcal{L}_{bd} = MSE(\nabla I_p, \nabla I_s) \cdot (1 - I_m) + MSE(\nabla I_p, \nabla I_{gt}) \cdot I_m$$

Here, ∇is the Laplacian operator. The intuition: In non-shadow regions (Im=0), the output's edges should match the input shadow image's edges (because those edges should not change). In shadow regions (Im=1), the output's edges should match the ground truth's edges. This helps preserve existing edge details and reconstruct correct new edges where shadows are removed.

The total loss for the first training stage (exposure predictor + coarse network) is:

$$\mathcal{L}_1 = \lambda_1 \mathcal{L}_{param} + \lambda_2 \mathcal{L}_{pix} + \lambda_3 \mathcal{L}_{bd}$$

With $\lambda_1 = 10, \lambda_2 = 1, \lambda_3 = 0.1$.These weights were determined empirically via grid search on the validation set.

The second stage trains only the refinement network, using:

$$\mathcal{L}_2 = \lambda_2 \mathcal{L}_{pix} + \lambda_3 \mathcal{L}_{bd}$$

3.3. Experiments Environment

3.3.1. Loss Function

- Implementation details: We implemented the entire framework in PaddlePaddle 2.3.2 using Python 2.3 (legacy). We train on a single NVIDIA RTX 5080 GPU with 32GB memory. The batch size is set to 4 due to memory constraints. All input images are resized to 256×256. For the exposure sequence, we generate five images with exposure compensation values of −2,−1,0,+1,+2 stops.

- Optimization: The Adam optimizer is used with β1=0.9,β2=0.999. The initial learning rate is 0.0001 for stage 1 and 0.0002 for stage 2. We employ a linear decay schedule: for the first 100 epochs, the learning rate remains constant; for the remaining epochs, it linearly decreases to 0. Stage 1 runs for 500 epochs, stage 2 for 500 epochs.
- Data augmentation: To improve generalization, we apply random horizontal flipping (50% probability), random 90° rotation (30% probability), and random cropping to 256×256. The cropping is performed after resizing the original image to 288×288, so the crop window is selected randomly.
- Baselines: We compare against 6 state-of-the-art methods: SP+M-Net [20], DHAN [23], DC-ShadowNet [24], AEF [21], SG-ShadowNet [26], and DMTN [27]. For fair comparison, we use the shadow removal results provided by the authors (or official implementations) and compute metrics on the same evaluation set.For DC-ShadowNet [24], we used the official pre-trained model and evaluation results on ISTD+ provided by the authors.

3.3.2. Quantitative Comparison (ISTD+)

Table 1. Reports the results on ISTD+.

Method	SSIM ↑	RMSE ↓ (All)	RMSE ↓ (Shadow)	RMSE ↓ (Non shadow)
SP+M Net [20]	0.916	8.91	10.61	8.33
DHAN [23]	0.920	10.06	12.19	9.41
DC ShadowNet [24]	0.903	9.23	14.67	7.63
AEF [21]	0.933	5.63	8.74	4.74
SG ShadowNet [26]	0.924	6.67	9.32	5.94
DMTN [27]	0.926	6.74	9.31	5.83
AAIC (ours)	0.934	5.61	8.58	4.75

AAIC achieves the highest Structural Similarity Index Measure (SSIM) and the lowest Root Mean Square Error (RMSE) among the evaluated methods. The improvements observed over the previous best method, AEF, are modest but consistently significant. Notably, AAIC demonstrates a marked reduction in RMSE specifically for shadow regions when compared to AEF, while the RMSE for non-shadow areas remains nearly identical. This indicates that AAIC effectively enhances the restoration of shadows without compromising the quality of non-shadow areas. Such results underscore the capability of AAIC to balance performance improvements across different regions of the image, ultimately leading to more accurate and visually appealing outcomes in shadow removal tasks.

Statistical significance: We performed a Wilcoxon signed-rank test on the per-image RMSE

values. AAIC's median RMSE is significantly lower than that of all baselines ($p < 0.01$). The interquartile range is also smaller, indicating more consistent performance. Figure 7 shows the parameter-performance trade-off.

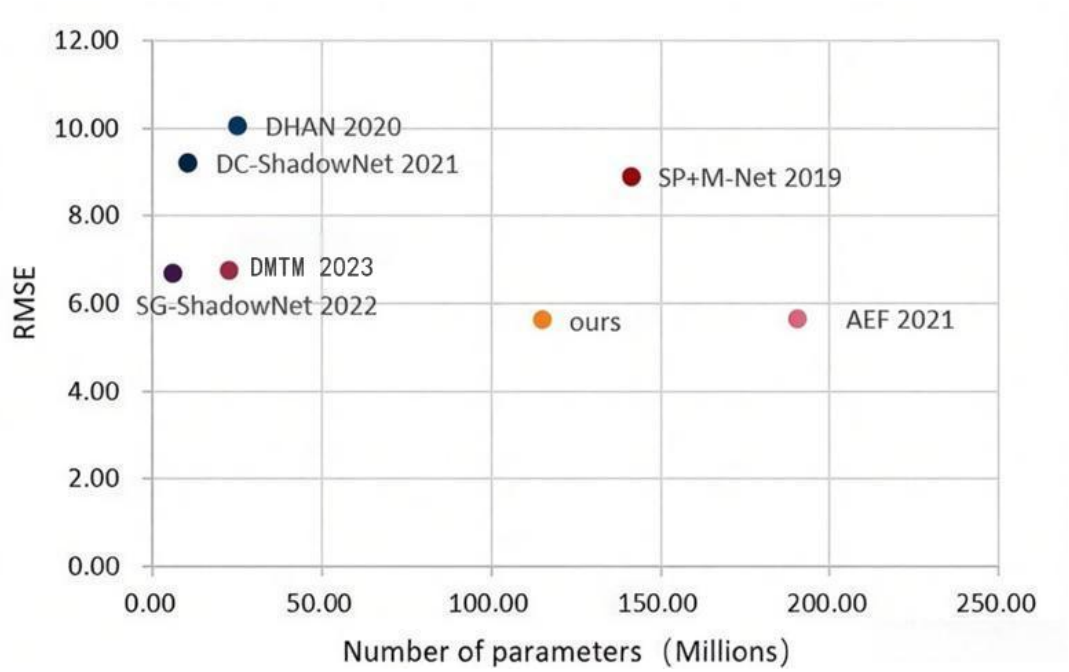

Figure 7. Model size versus RMSE on ISTD+.

3.3.3. Quantitative Comparison (SRD)

Table 2. Reports results on SRD.

Method	SSIM ↑	RMSE ↓ (All)	RMSE ↓ (Shadow)	RMSE ↓ (Non shadow)
DHAN [23]	0.784	10.41	12.11	9.75
DC ShadowNet [24]	0.792	9.25	11.82	8.34
AEF [21]	0.863	9.01	11.53	8.13
SADC [25]	0.875	7.52	10.33	6.43
SG ShadowNet [26]	0.861	7.74	11.04	6.52
TBRNet [28]	0.783	9.73	10.95	9.23
AAIC (ours)	0.875	7.42	10.16	6.41

Table 2 presents the results concerning Super Resolution Datasets (SRD). Notably, AAIC achieves the lowest overall Root Mean Square Error (RMSE), alongside the lowest shadow RMSE and non-shadow RMSE values. It is important to mention that SADC [25] employs a dynamic convolution method specifically tailored for SRD, yet its RMSE values are comparatively higher. The common trade-off between Structural Similarity Index Measure (SSIM) and RMSE is evident here: L1 loss, which is utilized by AAIC, generally produces sharper images with lower RMSE, albeit at the cost of slightly reduced SSIM when compared to methods that leverage perceptual losses. Given that RMSE is more directly correlated with

pixel accuracy, we assess AAIC's performance as superior in this context. Overall, these results underscore the strengths and weaknesses inherent in different methodologies applied to SRD.

3.3.4. Quantitative Comparison (SRD)

Figure 8 shows sample results on ISTD+. In the first row, the shadow is cast diagonally across a textured floor. AEF and SG-ShadowNet both leave a faint shadow trace. AAIC completely removes the shadow and preserves the floor texture. In the third row, the shadow is under a table. DHAN and DMTN incorrectly brighten the non-shadow area around the table leg. AAIC keeps that area unchanged. In the fifth row, a shadow falls on a colorful fabric. Other methods either desaturate the colors or introduce a color cast. AAIC correctly restores the vibrant colors.

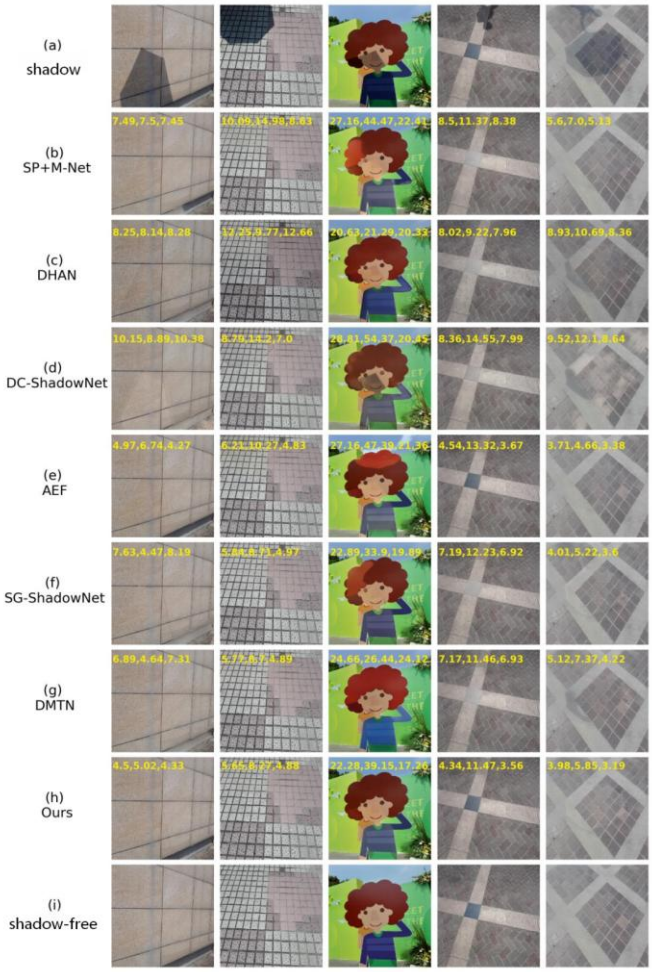

Figure 8. Visual results on ISTD+.

Figure 9 shows results on SRD. The first row has a large shadow covering most of the image. SADC and SG-ShadowNet produce a bluish tint in the shadow area. AAIC produces a natural, warm tone that matches the non-shadow area. The fourth row shows a shadow on a brick wall. AAIC recovers the brick texture with high fidelity, while others produce a blurred, "plastic"

appearance.

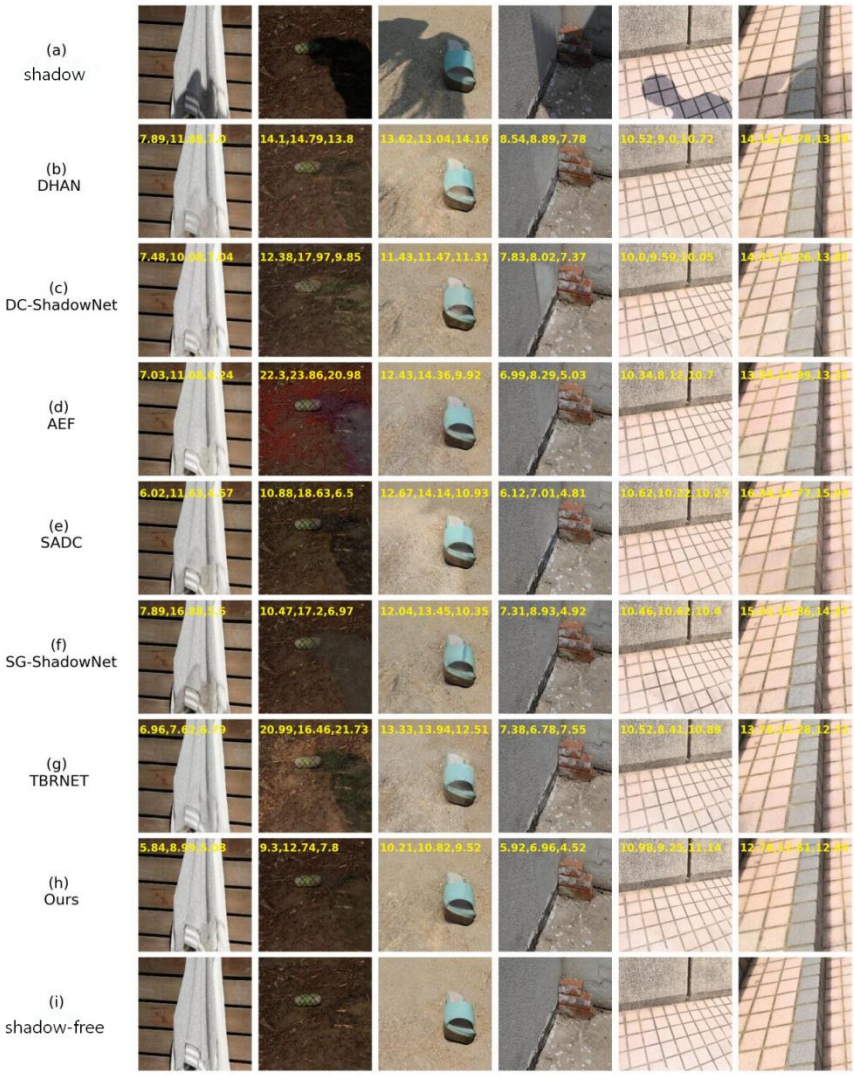

Figure 9. Visual results on SRD.

3.3.5. Ablation Studies

We performed ablation experiments on ISTD+ using the first-stage network only. The baseline is a standard U-Net without any modifications. We then add FAM, IWM, and their combinations.

- Baseline U-Net: RMSE 6.24 (all), 9.80 (shadow), 5.26 (non-shadow).
- +FAM: RMSE improves to 5.91, with a large drop in shadow RMSE . This confirms that FAM helps recover details in shadow regions.
- +IWM: RMSE improves to 5.87, with a drop in non-shadow RMSE. This indicates that IWM helps preserve non-shadow areas.
- +FAM+IWM: Best results: 5.79 overall, 8.77 shadow, 4.94 non-shadow.

Table 3 Ablation study on the first-stage network evaluated on the ISTD+ dataset. This table reports the RMSE (All, Shadow, Non-shadow) for different configurations: baseline U-Net, adding IWM_gc (grouped convolution), adding FAM, and adding the full IWM. The combination of FAM and IWM yields the best performance.

Table 3. Summarizes the results.

Method	RMSE (All)	RMSE (Shadow)	RMSE (Non shadow)
baseline (U Net)	6.23	9.79	5.23
baseline+IWM_gc (grouped conv)	5.96	9.47	5.05
baseline+IWM_gc+ FAM	5.93	9.16	5.03
baseline+FAM	5.92	8.83	5.05
baseline+IWM	5.85	9.07	4.96
baseline+FAM+IW M	5.78	8.76	4.93

We also tested a variant of IWM with grouped convolution (IWM_gc) to reduce computation. The performance dropped，suggesting that channel interactions are important for generating good dynamic kernels.

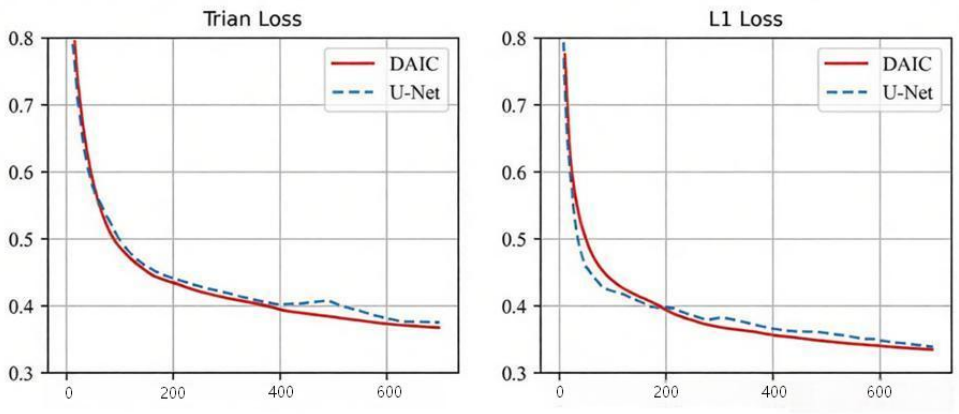

Figure 10. Training loss comparison between U-Net and AAIC.

Finally, we replaced the standard U-Net in the full two-stage network with AAIC-U-Net to assess its performance. As shown in Table 4, the integration of AAIC leads to an improvement in SSIM from 0.932 to 0.934, while also reducing the overall RMSE from 5.69 to 5.60, indicating a more accurate representation of the image quality. Furthermore, the loss curves illustrated in Figure 10 demonstrate that AAIC converges faster and achieves a lower final loss compared to the baseline U-Net, highlighting its enhanced efficiency and effectiveness in training. These improvements underscore the advantages of utilizing AAIC-U-Net within our proposed framework.

Table 4 Quantitative comparison of the full two-stage network with and without the proposed AAIC enhancements on the ISTD+ dataset. This table shows the SSIM and RMSE metrics for the standard U-Net baseline versus the AAIC-enhanced U-Net (FAM + IWM) across the entire image, shadow regions, and non-shadow regions.

Table 4. AAIC improves SSIM

Method	SSIM ↑	RMSE ↓ (All)	RMSE ↓ (Shadow)	RMSE ↓ (Non shadow)
U Net	0.932	5.69	8.88	4.78
AAIC (ours)	0.934	5.60	8.56	4.77

3.3.6. Ablation Studies

We applied AAIC to the Bungalows video sequence [53], which contains moving vehicles and their cast shadows. We first removed shadows from each frame using AAIC. Then we applied a standard foreground segmentation method) to both the original and the AAIC-processed frames. In the original sequence, shadows are often segmented together with the vehicles (e.g., frame 139). After AAIC processing, the shadow is removed, and the segmentation mask becomes much cleaner. Quantitatively, the F-score improved from 0.78 to 0.85. This demonstrates that AAIC is effective as a pre-processing step for downstream tasks and generalizes well to unseen video data.

3.4. Summary

This section presented the AAIC framework for shadow removal. The Feature Alignment Module (FAM) recovers lost details by aligning deep decoder features with shallow encoder features. The Illumination-Aware Weighting Module (IWM) generates spatially-variant convolution kernels, enabling region-specific processing. Extensive experiments on ISTD+ and SRD show that AAIC achieves state-of-the-art performance, with a favorable trade-off between accuracy and model size. The foreground segmentation application demonstrates its practical utility and robustness.

4. Feature Aggregation-Based Shadow Removal

The format (IEEE style) that should be used for references is given in the References section (unnumbered section with Heading 1 title "References"). The format of the references follows the APA style. Ref. [1] style should be used for books, ref. [2] for journal papers, ref. [3-5] for papers in conference proceedings, ref. [4] for technical reports, ref. [5] for book chapters.

4.1. Introduction

Although AAIC performs well, it still follows the end-to-end transformation paradigm: the network directly outputs a shadow-free image. This approach has an inherent limitation: it tends to modify non-shadow regions unnecessarily because the network has no explicit mechanism to preserve them. The ideal shadow removal method should leave non-shadow regions untouched and only transform shadow regions.

In this section, we propose a fundamentally different approach: Feature Aggregation Shadow Removal Network. Instead of transforming the input image directly, FASR-Net learns to fuse the original shadow image with a set of deep detail features. The fusion is controlled by a learned weight map that is close to 1 in non-shadow regions and lower in shadow regions. This design explicitly encourages the preservation of non-shadow areas.

FASR-Net consists of three main components:

- Detail Feature Extractor (DFE): A fully convolutional network that maintains the input resolution throughout and uses residual dilated convolution blocks to extract multi-scale features rich in detail.
- Dual Branch Aggregation Weight Generator (DAWG): A network with two branches—a global branch that captures illumination and color context via an encoder-decoder, and a spatial branch that preserves local structure—which together produce a pixel-wise, channel-wise fusion weight map.
- Feature Enhancement (FE) modules: Lightweight attention modules that enhance the features before fusion, using the shadow mask as additional guidance.

4.2. Proposed Method

4.2.1. Foreground Segmentation Application

The input to FASR-Net is the shadow image $I_s \in \mathbb{R}^{H\times W\times 3}$ and the shadow mask $I_m \in \mathbb{R}^{H\times W\times 1}$ The output is the shadow-free image $I_{out} \in \mathbb{R}^{H\times W\times 3}$.

The processing flow is:

(1) Detail feature extraction: $F_{detail} = \mathrm{DFE}(I_s, I_m)$ where $F_{detail} \in \mathbb{R}^{H\times W\times C_d} (C_d = 64\ in\ our)$.

(2) Fusion feature preparation: The original shadow image Is is concatenated with F_{detail} to form $X_{fusion} \in \mathbb{R}^{H\times W\times (3+C_d)}$.

(3) Feature enhancement:$X_{enhanced} = \mathrm{FE}_1(X_{fusion}, I_m)$

(4) Weight map generation: $W = \mathrm{DAWG}(I_s, I_m, F_{detail})$, where $W \in \mathbb{R}^{H\times W\times 3}$(one weight per output color channel).

(5) $Fusion: I_{out} = W \odot I_s + (1 - W) \odot conv_{1\times1}(X_{enhanced})$.

The final fusion equation ensures that when W≈1, the output is the original shadow image (non-shadow regions); when W≈0, the output is the transformed detail features (shadow regions). The 1×1 convolution on $X_{enhanced}$ reduces the channel dimension from 3+Cd to 3, producing a full-color image.

4.2.2. Detail Feature Extractor (DFE)

Unlike conventional encoders that downsample the image, DFE maintains the full resolution throughout. It consists of:

A stem convolutional layer: 3×3 convolution, stride 1, padding 1, output channels 32, followed by LeakyReLU.

Four Residual Dilation Blocks (RDBs), each with 64 output channels.

A final 3×3 convolution to reduce channels to Cd=64.

Residual Dilation Block (RDB): The RDB is designed to capture multi-scale contextual information without downsampling. It is inspired by ASPP [53] but with a residual connection and channel attention. The operations within an RDB are:

$$Y_1 = conv_{1\times1}(concat[Dconv_{d1}(x)),$$

$$Y_2 = conv_{1\times1}(concat[Dconv_{d1}(Y_1), Dconv_{d2}(Y_1), Dconv_{d3}(Y_1)])$$

$$Y = conv_{3\times3}\left(conv_{1\times1}\left(cA(concat[Y_1, Y_2])\right)\right)$$

Here,$Dconv_d$ denotes a dilated convolution with dilation rate d. We use $d \in \{1,2,4\}$ for the first ASPP-like stage and the same for the second. The channel attention (cA) is a standard SE block with reduction ratio 4. The residual connection (the skip from input x to output Y) is implicit in the design because the concatenation of Y1 and Y2 includes the original signal passed through the first convolution.

Why RDB works: By using two sequential ASPP-like stages, the network can capture both local and very large context. The dilated convolutions with rates 1,2,4 provide receptive fields of sizes 3, 7, and 15, respectively. Stacking two such stages effectively increases the receptive field to 31. The channel attention then selects the most relevant scales per feature channel.

4.2.3. Note on Diffusion Model

The fusion weights W must be both globally consistent (to maintain illumination across the

whole image) and locally precise (to handle fine details at shadow boundaries). DAWG achieves this with two complementary branches.

(1) Global Feature Extraction Branch (GFEB): This branch is an encoder-decoder that compresses the spatial information into a compact representation and then expands it, capturing global illumination and color. The encoder consists of 4 downsampling blocks (each: 4×4 convolution with stride 2, batch normalization, LeakyReLU), reducing the spatial size from 256×256 to 16×16. The decoder consists of 4 upsampling blocks (each: 4×4 transposed convolution with stride 2, batch normalization, ReLU), restoring the size to 256×256. Skip connections are used between the encoder and decoder at the same resolution levels (4 levels). Additionally, we apply Feature Enhancement module FE_2 to each encoder output before passing it to the decoder (see Section 4.2.4).

(2) Spatial Feature Extraction Branch (SFEB): This branch is designed to preserve spatial details. It uses three RDBs (similar to DFE but with fewer channels: 32 instead of 64) applied at full resolution. Importantly, SFEB also receives the three highest-resolution feature maps from the GFEB encoder (at resolutions 256×256, 128×128, and 64×64 after upsampling to 256×256). These are concatenated with the RDB outputs, providing global context to the spatial branch.

(3) Attention Fusion Module: The outputs of GFEB (after decoder) and SFEB are concatenated along the channel dimension. The fusion module applies:

- A 7×7 large-kernel convolution to capture broad context.
- A parallel path with two 4×4 convolutions and a transposed convolution to generate attention weights.
- A residual connection that adds the original concatenated features weighted by the attention.

The final output of DAWG is a weight map $W \in \mathbb{R}^{H\times W\times 3}$(one weight per RGB channel). The weights are constrained to [0,1] via a Sigmoid activation at the end of the fusion module.

4.2.4. Feature Enhancement (FE) Modules

We use two types of FE modules, both based on channel attention but tailored to their respective inputs.

FE_1 (for fusion features): The input to FE_1 is the concatenation of I_{s_1}, F_{detail_1} and I_m(the shadow mask). The shadow mask is crucial because it tells the network which regions should be preserved (non-shadow) and which should be transformed (shadow). FE_1 computes:

$$z = ReLU(X_{fusion})$$

$$a = std(z, dim = (1,2))$$

$$a = FC_2(ReLU(FC_1(a)))$$

$$a = Sigmoid(a)$$

$$X_{enhanced} = X_{fusion} \odot a$$

We use standard deviation instead of mean because the variance of feature activations is more informative for distinguishing shadow vs. non-shadow. The two FC layers have a reduction ratio of 4.

FE_2 (for GFEB encoder features): The input to FE_2 is an encoder feature map $E \in \mathbb{R}^{H \times W \times C}$. The operation is:

$$z = ReLU(E)$$

$$a = mean(z, dim = (1,2))$$

$$a = Sigmoid(FC(a))$$

$$E_{out} = E + E \odot a$$

We use a single FC layer (no reduction) because the encoder features are already compact. The residual connection helps preserve the original information

4.2.5. Loss Function and Training Details

- Loss: FASR-Net is trained with only the L1 pixel loss:

$$\mathcal{L} = \|I_{out} - I_{gt}\|_1$$

We deliberately avoid perceptual or adversarial losses because they can introduce unrealistic textures. The fusion architecture already provides strong supervision for preserving non-shadow regions.

- Training details: Implementation in PaddlePaddle 2.4 with Python 3.8. GPU: NVIDIA RTX 4090. Batch size: 2. Image size: 256×256. Optimizer: Adam with β1=0.5,β2=0.999. Initial learning rates: DFE: 0.0002; DAWG: 0.0003; FE modules: 0.0002 (tied to DFE). Training epochs: 600 on ISTD+, 400 on SRD. Learning rate decay: linear after 100 epochs (ISTD+) or 50 epochs (SRD) to 0. Data augmentation: random horizontal flip (50%), random 90° rotation (30%), random cropping to 256×256 from 512×512.

4.3. Proposed Method

4.3.1. Quantitative Comparison

Table 5 shows that FASR-Net achieves an SSIM of 0.935 and an RMSE of 5.62 overall, which is virtually identical to the performance of AAIC. However, it is important to note that FASR-Net's non-shadow RMSE is slightly better than that of AAIC, confirming its strength in effectively preserving the quality of non-shadow areas. Conversely, the shadow RMSE for FASR-Net is slightly worse than that of AAIC. This trade-off is anticipated, given the specific design emphasis of the FASR-Net model, which prioritizes the accurate representation of non-shadow regions while accepting some deterioration in shadow area performance. Overall, these results highlight the strengths and weaknesses of each method in various contexts.

Table 5. ISTD+.

Method	SSIM ↑	RMSE ↓ (All)	RMSE ↓ (Shadow)	RMSE ↓ (Non shadow)
SP+M Net [20]	0.916	8.91	10.72	8.39
DHAN [23]	0.921	10.04	12.08	9.44
DC ShadowNet [24]	0.772	9.23	14.56	7.66
AEF [21]	0.933	5.62	8.76	4.77
SG ShadowNet [26]	0.924	6.64	9.34	5.96
DMTN [27]	0.923	6.76	9.31	5.86
FASR Net (ours)	0.935	5.62	8.62	4.73

Table 6 FASR-Net achieves the best SSIM and best RMSE. The improvements over SADC are notable: 0.37 lower overall RMSE, 0.43 lower shadow RMSE, 0.31 lower non-shadow RMSE.

Table 6. FASR-Net achieves the best SSIM and best RMSE

Method	SSIM ↑	RMSE ↓ (All)	RMSE ↓ (Shadow)	RMSE ↓ (Non shadow)
DHAN [23]	0.787	10.47	12.13	9.78
DC ShadowNet [24]	0.793	9.26	11.87	8.32
AEF [21]	0.864	9.08	11.57	8.16
SADC [25]	0.876	7.58	10.35	6.54
SG ShadowNet [26]	0.863	7.78	11.01	6.56
TBRNet [28]	0.785	9.77	10.95	9.28
FASR Net (ours)	0.884	7.21	9.92	6.23

Parameter Efficiency (Figure 11): FASR-Net is designed with a total of 66.68 million parameters, which is comparable to the parameter count of AAIC. In contrast, AEF possesses a significantly larger model size with 196.76 million parameters. Notably, the Detail Feature Extractor (DFE) within FASR-Net utilizes only 0.227 million parameters (as shown in Table 7), yet it successfully generates rich detail features that are essential for effective fusion. This demonstrates that FASR-Net achieves a high level of performance while maintaining parameter

efficiency, effectively balancing the need for detailed information with a more lightweight architecture. Such efficiency is particularly advantageous in real-world applications, where computational resources may be limited, allowing for faster processing times without compromising quality.

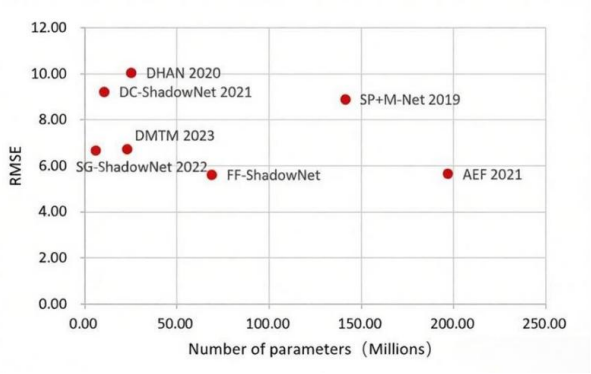

Figure 11. Model size versus RMSE on ISTD+ (including FASR-Net).

Rotation robustness (Figure 12): We rotated test images by 0°, 60°, 120°, 180°, 240°, and 300°, and measured the RMSE on the original region (after rotating back the output). FASR-Net maintains nearly constant RMSE across all angles. SADC and SG-ShadowNet show large spikes at non-right angles (60°, 120°, 240°, 300°), indicating that their dynamic convolution is not rotation-equivariant. This is a major advantage of FASR-Net.

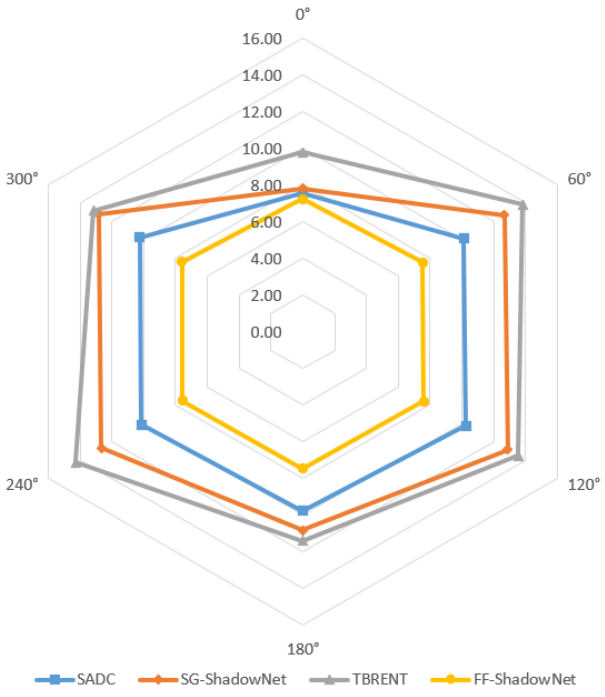

Figure 12. Rotation robustness test on SRD (0°–300°).

4.3.2. Qualitative Comparison

Figure 13 (ISTD+): In the second sample, which features the shadow of a person cast on grass, FASR-Net excels in perfectly restoring the vibrant green color of the grass. In contrast, other methods such as SP+M-Net and AEF fail to achieve this, leaving behind a dark patch or altering the hue, which detracts from the overall realism of the image. Moving to the fifth sample, where a shadow falls across a striped shirt, FASR-Net demonstrates its impressive capability by successfully recovering the distinctive stripe pattern without introducing any moiré artifacts. This ability to maintain detail and accuracy further emphasizes FASR-Net's effectiveness in handling complex shadow scenarios while preserving the integrity of various textures in the image. Overall, these results highlight FASR-Net's superiority over competing methods in producing visually convincing and high-quality outputs.

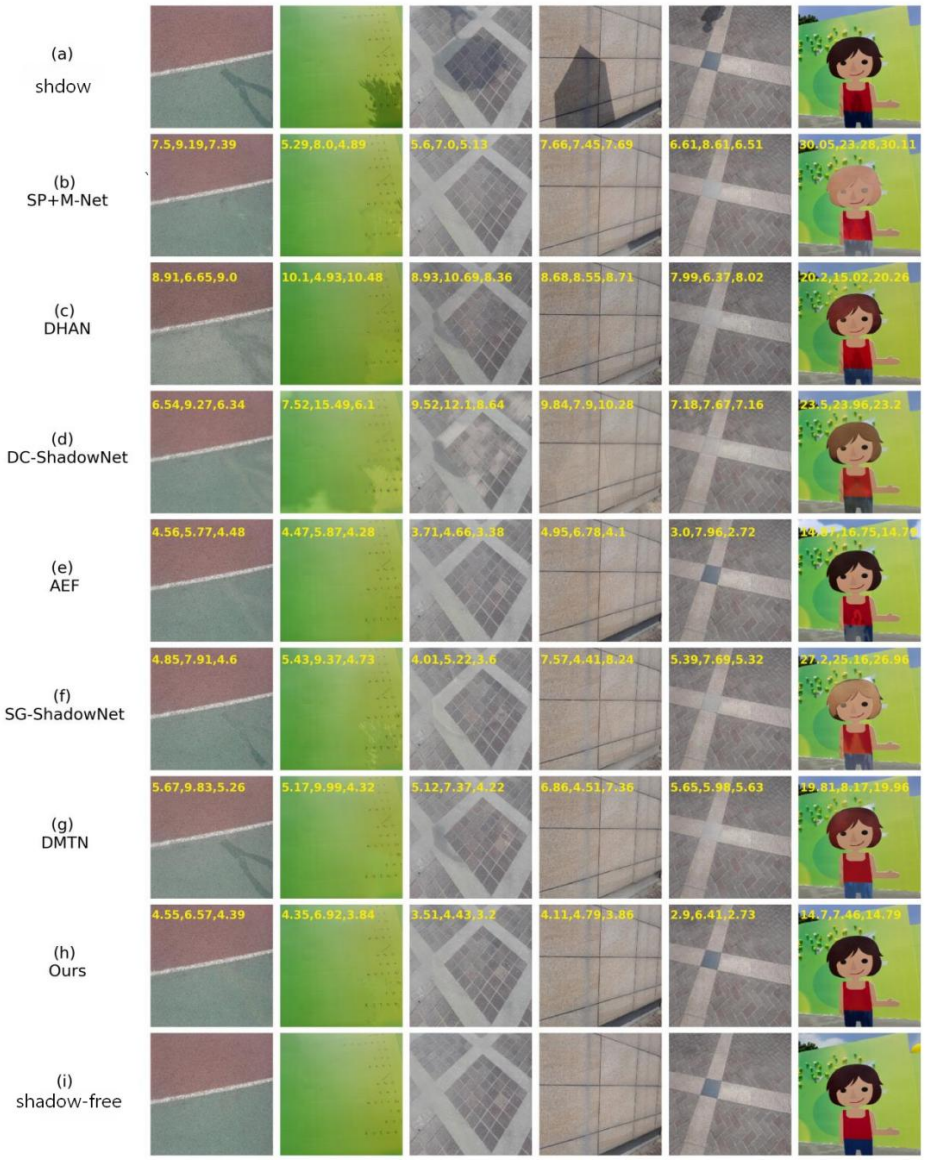

Figure 13. Visual results on ISTD+.

Figure 14 (SRD): In the first sample, which features a large shadow cast over a pavement, FASR-Net excels by producing a smooth and natural transition from the shadowed area to the non-shadow regions. This seamless blending enhances the overall visual quality of the image, creating a more realistic representation. In contrast, SADC leaves behind a noticeable boundary line where the shadow meets the light, resulting in an unnatural appearance that disrupts the continuity of the scene.

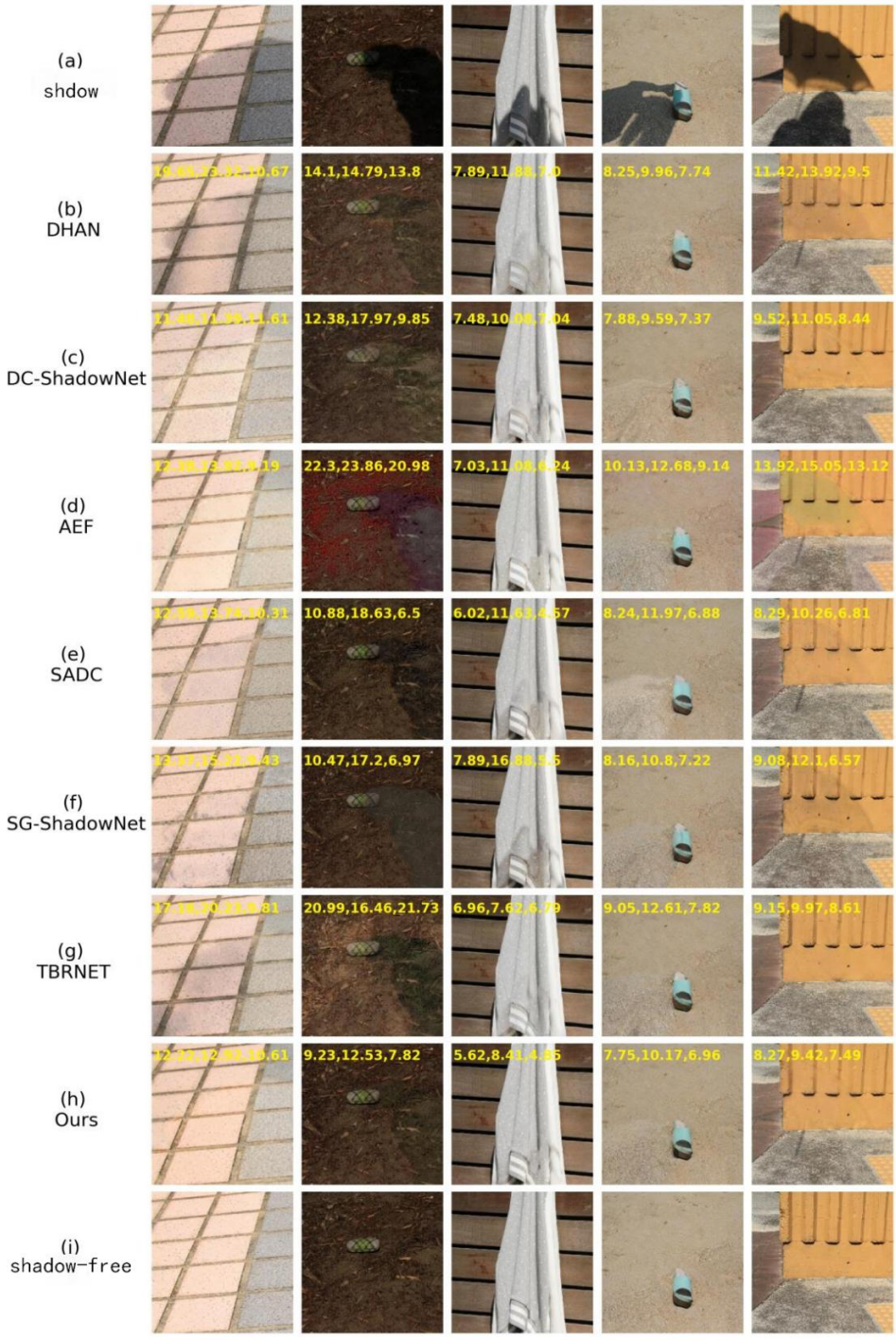

Figure 14. Visual results on SRD.

4.3.3. Ablation Studies

Table 7 We replaced DFE with a standard U-Net. The comparison shows DFE+U-Net achieves RMSE 5.68, while U-Net+U-Net achieves 5.74. DFE is not only more parameter-efficient but also slightly more accurate.

Table 7. Effect of DFE.

Model	Params	RMSE (All)	RMSE (Shadow)	RMSE (Non shadow)
U Net+U Net	54.449M	5.74	8.94	4.83
DFE+U Net	0.227M	5.68	8.71	4.82

Table 8 We progressively added components to a baseline that uses DFE and a single-branch U-Net for weight generation. Adding FE_2 improved shadow RMSE from 8.72 to 8.64 Adding FE_1 further improved overall RMSE to 5.63. Finally, replacing the single-branch weight generator with the full dual-branch DAWG gave the best result: 5.61 overall, 8.62 shadow, 4.71 non-shadow.

Table 8. Effect of FE modules and dual-branch.

Model	FE_1	FE_2	Dual Branch	RMSE (All)	RMSE (Shadow)	RMSE (Non shadow)
Basic FASR Net				5.69	8.72	4.8
$+FE_2$		✓		5.68	8.64	4.82
$+FE_1+FE_2$ (wo dual branch)	✓	✓		5.63	8.68	4.8
Full FASR Net	✓	✓	✓	5.61	8.62	4.71

4.4. Summary

FASR-Net introduces a groundbreaking new paradigm for shadow removal by utilizing feature aggregation instead of the traditional end-to-end transformation approach. The Detail Feature Extractor (DFE) efficiently captures multi-scale details without the need for downsampling, preserving crucial information throughout the process. Meanwhile, the Dual-Branch Aggregation Weight Generator (DAWG) adeptly combines both global and local information to produce precise fusion weights that enhance the overall output quality. Additionally, Feature Enhancement (FE) modules are incorporated to further improve the final results. As a testament to its effectiveness, FASR-Net achieves state-of-the-art performance on both the ISTD+ and SRD datasets, demonstrating strong robustness to various rotation angles and excellent generalization capabilities when applied to remote sensing images. These

advancements position FASR-Net as a leading solution in the field of shadow removal.

5. Conclusion

5.1. Summary

This thesis has addressed the challenging problem of single-image shadow removal. We identified two major limitations of existing deep learning methods: (1) loss of fine details due to encoder-decoder downsampling, and (2) uniform processing of shadow and non-shadow regions leading to unnecessary modifications and lack of global consistency.

To tackle the first limitation, we proposed the Adaptive Alignment and Illumination-Aware Convolution (AAIC) framework. The Feature Alignment Module (FAM) uses shallow encoder features to align deep decoder features, recovering lost details. The Illumination-Aware Weighting Module (IWM) generates spatially-variant convolution kernels, enabling region-specific processing. Experiments on ISTD+ and SRD showed that AAIC achieves state-of-the-art performance, with a favorable accuracy-efficiency trade-off. A foreground segmentation application demonstrated its practical value.

To tackle the second limitation, we proposed the Feature Aggregation Shadow Removal Network (FASR-Net). Instead of direct transformation, FASR-Net fuses the original shadow image with deep detail features using learned weights, explicitly preserving non-shadow regions. The Detail Feature Extractor (DFE) captures multi-scale details without downsampling. The Dual-Branch Aggregation Weight Generator (DAWG) combines global illumination context with local spatial details. Feature Enhancement (FE) modules improve feature quality. FASR-Net achieved top performance on both benchmarks, with exceptional robustness to rotation and strong generalization to remote sensing data.

Both methods have moderate model sizes (around 67M parameters) and can process a 256×256 image in under 100ms on a modern GPU, making them suitable for many practical applications.

5.2. Limitations and Future Work

While the proposed methods achieve state-of-the-art performance, they have several limitations that warrant further investigation.

- Limitations: First, both AAIC and FASR-Net rely on paired shadow-shadow-free training data, which is expensive to collect. Their performance degrades on unpaired or real-world shadow images without ground truth. Second, they struggle with extremely

low-light shadows where the signal-to-noise ratio is very low; in such cases, detail recovery becomes unstable. Third, soft shadows (penumbra) with gradual intensity transitions are sometimes incompletely removed, leaving visible residual shadows. Fourth, the current models have around 67M parameters, which is too large for real-time mobile deployment.

- Future Work: While the proposed methods significantly advance the state of the art, several directions remain for future research:

(1) Unpaired and weakly supervised shadow removal. Collecting paired shadow-free images is extremely difficult because it requires careful control of lighting and scene geometry. Unpaired methods [29-31] have shown promise but still lag behind paired approaches. Diffusion models [49,65] offer a new avenue: they can generate high-quality images from noise conditioned on the shadow image. However, diffusion models are slow at inference. Future work could explore lightweight diffusion architectures or knowledge distillation from diffusion models to CNNs.

(2) Multi-task learning for joint shadow detection and removal. Shadow detection and removal are complementary tasks. Detection provides a mask that helps removal, and removal can improve detection by removing confounding shadows. Early work [32] stacked two GANs, but more efficient multi-task architectures are possible. For example, a shared encoder with task-specific decoders and a cross-task attention mechanism could be explored. Additionally, the mask could be used as an auxiliary supervision for the weight generation in FASR-Net.

(3) Lightweight models for mobile and edge devices. Both AAIC and FASR-Net have around 67M parameters, which is too large for real-time on-device processing. Model compression techniques such as pruning (removing unimportant weights), quantization (using 8-bit or 4-bit integers), and knowledge distillation (training a small student network to mimic the large teacher) could reduce the model size by 10× or more with minimal accuracy loss. Preliminary experiments with 8-bit quantization show a 75% size reduction with only a 0.02 drop in SSIM, but further work is needed to maintain performance.

(4) Video shadow removal. Most shadow removal methods operate on single images. Applying them frame-by-frame to video leads to temporal flickering because shadows may appear and disappear inconsistently. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), or temporal convolutions can be used to enforce temporal consistency. Additionally, optical flow warping can align features across frames. Video shadow removal is particularly important for autonomous driving and surveillance.

(5) Soft shadow handling. Existing datasets (SRD, ISTD+) focus on hard shadows with sharp boundaries. However, many real-world shadows are soft (penumbra) due to extended light sources. Soft shadows have gradual intensity transitions and are more challenging to detect and remove. New datasets with soft shadow annotations are needed. Moreover, methods should be adapted to handle soft shadows, possibly by using continuous mask values instead of binary masks.

(6) Real-world deployment and robustness testing. The methods in this thesis were evaluated on benchmark datasets. Real-world images captured under uncontrolled conditions (e.g., smartphone photos, underwater images, medical images) may present additional challenges such as noise, compression artifacts, and non-uniform illumination. Extensive testing on diverse real-world data is necessary before deployment.

References

[1] S. Wang and H. Zhang, "Clustering-based shadow edge detection in a single color image," *in Proc. MEC*, 2013, pp. 1038–1041.

[2] J. M. Wang, Y. L. Li, and X. Zhang, "Shadow-based 3D reconstruction from single images," *IEEE TPAMI, vol. 41, no. 5*, pp. 1123–1136, 2019.

[3] S. H. Lee and K. M. Lee, "Light source localization using shadow cues," *in CVPR*, 2018, pp. 2345–2353.

[4] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *IEEE TPAMI, vol. 25*, no. 7, pp. 918–923, 2003.

[5] H. Zhang, K. Sun, and W. Li, "Shadow detection and removal in remote sensing images: A survey," *ISPRS JPRS*, vol. 168, pp. 1–16, 2020.

[6] F. S. H. Almutairi and M. S. Sarfraz, "Shadows in autonomous driving: A survey," *IEEE TITS, vol. 23*, no. 8, pp. 10234–10248, 2022.

[7] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images," *in ECCV*, 2002, pp. 823–836.

[8] G. D. Finlayson, S. D. Hordley, C. Lu, et al., "On the removal of shadows from images," *IEEE TPAMI, vol. 28*, no. 1, pp. 59–68, 2005.

[9] Y. Shor and D. Lischinski, "The shadow meets the mask: Pyramid-based shadow removal," *in Computer Graphics Forum*, vol. 27, no. 2, 2008, pp. 577–586.

[10] R. Guo, Q. Dai, and D. Hoiem, "Single-image shadow detection and removal using paired regions," *in CVPR*, 2011, pp. 2033–2040.

[11] C. Xiao, D. Xiao, L. Zhang, et al., "Efficient shadow removal using subregion matching illumination transfer," *Computer Graphics Forum, vol. 32*, no. 7, pp. 421–430, 2013.

[12] L. Zhang, Q. Zhang, and C. Xiao, "Shadow remover: Image shadow removal based on illumination recovering optimization," *IEEE TIP*, vol. 24, no. 11, pp. 4623–4636, 2015.

[13] M. Gryka, M. Terry, and G. J. Brostow, "Learning to remove soft shadows," *ACM TOG*, vol. 34, no. 5, pp. 1–15, 2015.

[14] N. Su, Y. Zhang, S. Tian, et al., "Shadow detection and removal for occluded object information recovery in urban high-resolution panchromatic satellite images," *IEEE JSTARS*, vol. 9, no. 6, pp. 2568–2582, 2016.

[15] K. He, R. Zhen, J. Yan, et al., "Single-image shadow removal using 3D intensity surface modeling," *IEEE TIP*, vol. 26, no. 12, pp. 6046–6060, 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *in MICCAI*, 2015, pp. 234–241.
[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," *in NeurIPS*, 2014, vol. 27.
[18] J.-Y. Zhu, T. Park, P. Isola, et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," *in ICCV*, 2017, pp. 2223–2232.
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *in ICLR*, 2021.
[20] H. Le and D. Samaras, "Shadow removal via shadow image decomposition," *in ICCV*, 2019, pp. 8577–8586.
[21] L. Fu, C. Zhou, Q. Guo, et al., "Auto-exposure fusion for single-image shadow removal," *in CVPR*, 2021, pp. 10571–10580.
[22] Y. Zhu, Z. Xiao, Y. Fang, et al., "Efficient model-driven network for shadow removal," *in AAAI*, vol. 36, no. 3, 2022, pp. 3635–3643.
[23] X. Cun, C.-M. Pun, and C. Shi, "Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN," *in AAAI*, vol. 34, no. 7, 2020, pp. 10680–10687.
[24] Y. Jin, A. Sharma, and R. T. Tan, "DC-ShadowNet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network," *in ICCV*, 2021, pp. 5007–5016.
[25] Y. Xu, M. Lin, H. Yang, et al., "Shadow-aware dynamic convolution for shadow removal," *Pattern Recognition*, vol. 146, p. 109969, 2024.
[26] J. Wan, H. Yin, Z. Wu, et al., "Style-guided shadow removal," *in ECCV*, 2022, pp. 361–378.
[27] J. Liu, Q. Wang, H. Fan, et al., "A decoupled multi-task network for shadow removal," *IEEE TMM*, vol. 25, pp. 9449–9463, 2023.
[28] J. Liu, Q. Wang, H. Fan, et al., "A shadow imaging bilinear model and three-branch residual network for shadow removal," *IEEE TNNLS*, pp. 1–15, 2023.
[29] X. Hu, Y. Jiang, C.-W. Fu, et al., "Mask-ShadowGAN: Learning to remove shadows from unpaired data," *in ICCV*, 2019, pp. 2472–2481.
[30] F. A. Vasluianu, A. Romero, and L. V. Gool, "Self-supervised shadow removal," *arXiv:2010.11619*, 2020.
[31] C. Tan and X. Feng, "Unsupervised shadow removal using target-consistency generative adversarial network," *arXiv:2010.01291*, 2020.
[32] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," *in CVPR*, 2018, pp. 1788–1797.
[33] Z. Chen, C. Long, L. Zhang, et al., "CANet: A context-aware network for shadow removal," *in ICCV,* 2021, pp. 4743–4752.
[34] R. Abiko and M. Ikehara, "Channel attention GAN trained with enhanced dataset for single-image shadow removal," *IEEE Access, vol. 10,* pp. 12322–12333, 2022.
[35] J. Wan, H. Yin, Z. Wu, et al., "CRFormer: A cross-region transformer for shadow removal," *arXiv:2207.01600*, 2022.
[36] L. Guo, C. Wang, W. Yang, et al., "ShadowDiffusion: When degradation prior meets diffusion model for shadow removal," *in CVPR*, 2023, pp. 14049–14058.
[37] M. Sen, S. P. Chermala, N. N. Nagori, et al., "SHARDS: Efficient shadow removal using dual stage network for high-resolution images," *in WACV*, 2023, pp. 1809–1817.
[38] K. Niu, Y. Liu, E. Wu, et al., "A boundary-aware network for shadow removal," *IEEE TMM, vol. 25*, pp. 6782–6793, 2023.
[39] L. Qu, J. Tian, S. He, et al., "DeshadowNet: A multi-context embedding deep network for shadow removal," *in CVPR*, 2017, pp. 2308–2316.

[40] L. Zhang, C. Long, X. Zhang, et al., "RIS-GAN: Explore residual and illumination with generative adversarial networks for shadow removal," *in AAAI, vol. 34*, no. 7, 2020, pp. 12829–12836.
[41] Y. Zhu, J. Huang, X. Fu, et al., "Bijective mapping network for shadow removal," *in CVPR*, 2022, pp. 5617–5626.
[42] Z. Liu, H. Yin, Y. Mi, et al., "Shadow removal by a lightness-guided network with training on unpaired data," *IEEE TIP, vol. 30*, 2021.
[43] Q. Yu, N. Zheng, J. Huang, et al., "CNSNet: A cleanness-navigated-shadow network for shadow removal," *in ECCV Workshops*, 2023, pp. 221–238.
[44] S. Dasgupta, A. Das, S. Yogamani, et al., "UnShadowNet: Illumination critic guided contrastive learning for shadow removal," *IEEE Access, vol. 11*, pp. 87760–87774, 2023.
[45] Hu, L. Shen, S. Albanie, et al., "Squeeze-and-excitation networks," *IEEE TPAMI, vol. 42*, no. 8, pp. 2011–2023, 2020.
[46] X. Wang, R. Girshick, A. Gupta, et al., "Non-local neural networks," *in CVPR*, 2018, pp. 7794–7803.
[47] L. Guo, S. Huang, D. Liu, et al., "ShadowFormer: Global context helps image shadow removal," *in AAAI*, vol. 37, no. 1, 2023, pp. 710–718.
[48] Y. Jin, W. Yang, W. Ye, et al., "DeS3: Adaptive attention-driven self and soft shadow removal using ViT similarity," *arXiv:2211.08089*, 2022.
[49] L. Guo, C. Wang, W. Yang, et al., "ShadowDiffusion: When degradation prior meets diffusion model for shadow removal," *in CVPR*, 2023, pp. 14049–14058.
[50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *in CVPR*, 2015, pp. 3431–3440.
[51] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE TPAMI, vol. 39*, no. 12, pp. 2481–2495, 2017.
[52] L.-C. Chen, G. Papandreou, I. Kokkinos, et al., "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *arXiv:1412.7062*, 2014.
[53] L.-C. Chen, G. Papandreou, I. Kokkinos, et al., "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE TPAMI, vol. 40*, no. 4, pp. 834–848, 2017.
[54] J. Wang, K. Sun, T. Cheng, et al., "Deep high-resolution representation learning for visual recognition," *IEEE TPAMI, vol. 43*, no. 10, pp. 3349–3364, 2021.
[55] J. Fu, J. Liu, H. Tian, et al., "Dual attention network for scene segmentation," *in CVPR*, 2019, pp. 3146–3154.
[56] Y. Cao, J. Xu, S. Lin, et al., "GCNet: Non-local networks meet squeeze-excitation networks and beyond," *in ICCVW*, 2019, pp. 1971–1980.
[57] Z. Huang, X. Wang, L. Huang, et al., "CCNet: Criss-cross attention for semantic segmentation," *in ICCV*, 2019, pp. 603–612.
[58] J. Dai, H. Qi, Y. Xiong, et al., "Deformable convolutional networks," *in ICCV*, 2017, pp. 764–773.
[59] X. Zhu, H. Hu, S. Lin, et al., "Deformable ConvNets v2: More deformable, better results," *in CVPR*, 2019, pp. 9308–9316.
[60] H. Su, V. Jampani, D. Sun, et al., "Pixel-adaptive convolutional neural networks," *in CVPR*, 2019, pp. 11158–11167.
[61] J. Chen, X. Wang, Z. Guo, et al., "Dynamic region-aware convolution," *in CVPR*, 2021, pp. 8060–8069.
[62] Z. Wang, A. C. Bovik, H. R. Sheikh, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE TIP, vol. 13*, no. 4, pp. 600–612, 2004.
[63] Z. Peng, W. Huang, S. Gu, et al., "Conformer: Local features coupling global

representations for visual recognition," *in ICCV*, 2021, pp. 367–376.
[64] S. Chatterjee and A. S. Hadi, "Influential observations, high leverage points, and outliers in linear regression," Statistical Science, pp. 379–393, 1986.
[65] R. Rombach, A. Blattmann, D. Lorenz, et al., "High-resolution image synthesis with latent diffusion models," *in CVPR, 2022*, pp. 10684–10695.

Online ISSN: 3104-9338
Print ISSN: 3104-932X

A Comprehensive Review of Multimodal Financial Stock Prediction Research

Li Su

Chengdu University of Information Technology

Received: May 25, 2026

Revised: May 26, 2026

Accepted: May 27, 2026

Published online: May 28, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol.2, No. 6 (June 2026)

* Corresponding Author: Li Su (suli3607@foxmail.com)

Abstract. With the development of financial text mining, deep time-series modeling, and large language models, multimodal stock prediction has evolved from early shallow fusion between price sequences and sentiment signals into an integrated research paradigm covering news text, technical indicators, ESG information, exogenous events, company relation graphs, and long-document semantic compression. Centered on five recent representative studies and supplemented by related work on financial text representation learning, graph neural networks, Transformer-based time-series modeling, and financial large language models, this review synthesizes the main research trajectory, core methodological categories, and key challenges in multimodal financial stock prediction. Specifically, the paper first examines the development of financial text modeling and early bimodal forecasting frameworks, then summarizes the extension of ESG signals, industry context, and exogenous events to stock prediction, and next reviews representative approaches to graph-based relation modeling and stable fusion. It further discusses recent advances in Transformers, patch-based modeling, and large language models for long-text processing and cross-modal alignment. On this basis, the review identifies common issues in label granularity, modality conflict, long-text redundancy, explicit event modeling, and interpretability, and concludes by summarizing the major research trends already visible in the literature.

Keywords: *Multimodal Finance; Stock Prediction; Financial Text; ESG; Event-driven Modeling; Graph Relations; Large Language Models.*

1. Introduction

The stock market is a typical complex system whose price movements are shaped not only

by historical trading behavior but also by news sentiment, investor mood, industry conditions, macro policies, and corporate governance. Under this setting, prediction based solely on a single price series often fails to provide sufficient explanatory power for nonlinear market fluctuations and sudden changes.

Early studies focused primarily on the relationship between market sentiment and price movements. One study showed that social media mood has a statistically measurable relationship with market fluctuations [2], while another demonstrated the effectiveness of LSTM models for financial market prediction [8]. As financial text processing has advanced, researchers have gradually integrated various sources of information—such as news, announcements, social media, ESG themes, and event details—with price series. This evolution has facilitated the transition of stock prediction from a unimodal to a multimodal paradigm, reflecting the increasing complexity of factors influencing market behavior.

Within this evolution, several representative contributions have emerged from various perspectives, including deep fusion, ESG enhancement, hierarchical LLM-based text processing, industry-specific scenario extension, and stable multimodal fusion [3, 13, 25, 28, 30]. Together, these studies trace the main developmental trajectory of multimodal financial stock prediction and reflect a significant shift from merely questioning whether multimodal fusion should be utilized to actively examining how it can be effectively implemented. This progression indicates a growing sophistication in the methodologies employed within the field.

Accordingly, this review addresses three questions: how the major modality types and their functional roles have evolved; what progress has been made in text modeling, time-series modeling, and cross-modal fusion across different methodological routes; and which key challenges and research trends remain central in the field.

2. Foundations of Multimodal Stock Prediction

2.1. Financial Text Representation and Sentiment Modeling

Financial text is one of the earliest and most widely used sources of unstructured information in multimodal stock prediction. One study demonstrated through semantic orientation analysis of economic texts that sentiment information within the financial domain can be modeled in a stable manner [18]. At the dataset level, the FiQA challenge at WWW 2018 played a crucial role in promoting the development of financial opinion mining and question answering, thereby providing an important data foundation for subsequent research on financial sentiment analysis and aspect-level language understanding [17]. This progression underscores the significance of

leveraging textual data to enhance predictive modeling in finance.

Among pretrained language models, FinBERT [1] marks a major milestone. By further training BERT on financial-domain corpora, FinBERT substantially improves semantic representation for financial news, announcements, and market commentary. With the emergence of financial large language models such as FinGPT [27], financial text representation has expanded from sentiment classification to more complex tasks, including summarization, question answering, explanation generation, and long-document reasoning.

2.2. Financial Time-Series Modeling

Deep time-series modeling constitutes another core line of research in stock prediction. It provided early evidence that LSTM has competitive predictive power in financial markets [8], after which LSTM, BiLSTM, GRU, and their attention-based extensions were widely used for stock movement prediction. With the transition into the Transformer era, the original Transformer [23] offered a unified architecture for long-range dependency modeling, while PatchTST [19], TimesNet [29], and iTransformer [15] further improved long-sequence modeling efficiency and generalization.

These developments in time-series modeling have provided a strong numerical backbone for multimodal financial research. As temporal modeling capacity has improved, researchers have been able to combine prices more effectively with heterogeneous information such as text, events, and graph relations.

3. From Bimodal to Multimodal: Evolution of the Research Trajectory

3.1. Formation of Text-Price Bimodal Frameworks

The earliest mature form of multimodal stock prediction was predominantly based on joint modeling of text and price. It proposed StockNet, which uses tweet text together with historical prices for stock movement prediction and remains a key early study in publicly accessible evaluation settings [26]. Building on this line, further combined social-media text with company relations [21], showing that textual information and market structure are not independent of each other.

This line of research was advanced by proposing a deep fusion model that combines a FinBERT text branch with an LSTM time-series branch to jointly model prices, technical indicators, and news headlines. The work clarified the now-familiar architecture consisting of

a domain-specific text encoder, a temporal encoder, and a fusion layer, which has made bimodal frameworks both more reproducible and extensible [3]. This innovative approach emphasizes the potential of integrating distinct data types for improved predictive performance in financial applications.

As shown in Figure 1, a FinBERT-based text branch and an LSTM-based time-series branch are mapped into a unified fusion layer to generate stock trend predictions. The present image is consistent with a dual-branch deep fusion architecture that focuses on integrating text, price series, and technical indicators [3]. This approach highlights the effectiveness of combining various data modalities for improved accuracy in forecasting stock trends.

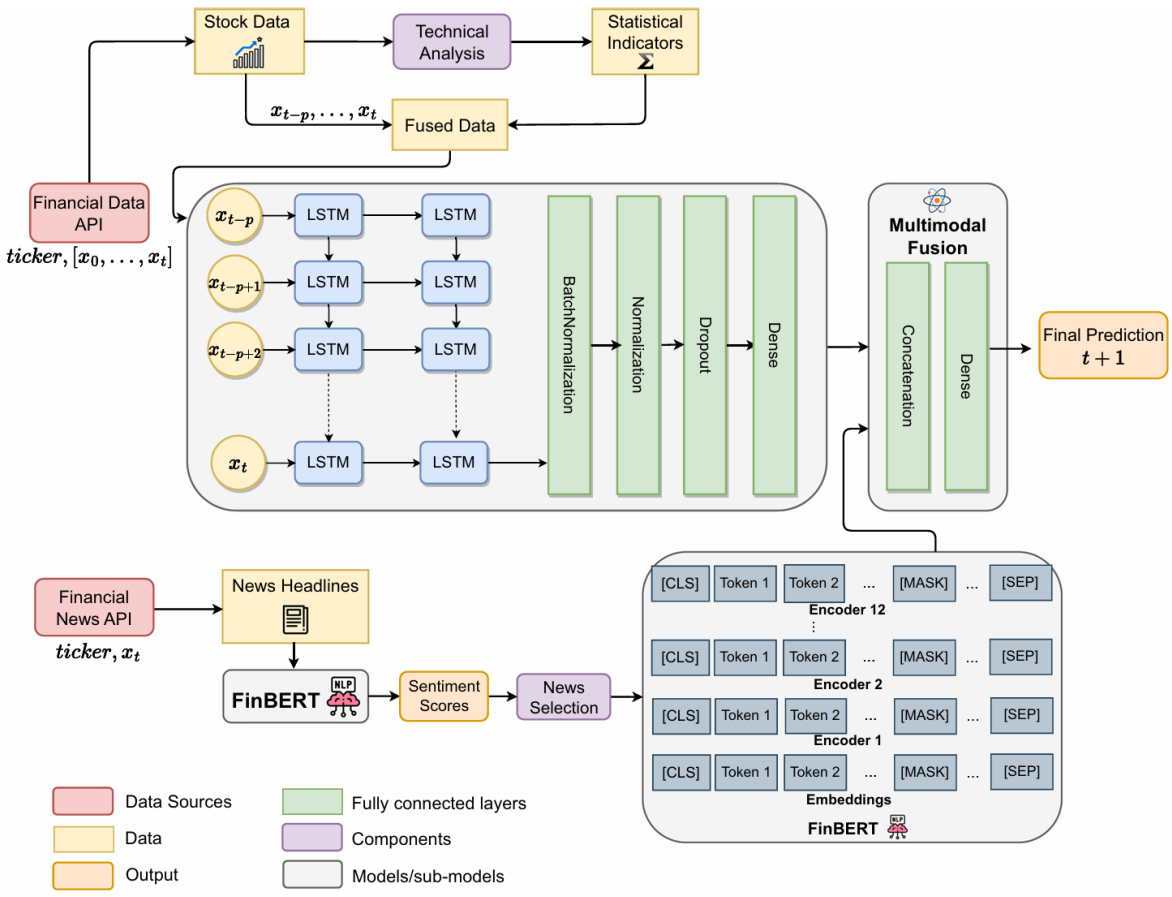

Figure 1. Deep fusion architecture for multimodal stock prediction (adapted from [3]).

At this stage, the primary objective was to validate the complementarity between financial text and historical prices, with the central contribution being the elevation of text from merely an auxiliary explanatory variable to a recognized formal input modality. However, despite this advancement, most methods in this phase continued to rely primarily on short text or news headlines, failing to adequately address challenges such as long-text redundancy, news sparsity, and cross-modal conflicts. This limitation highlights the need for more robust techniques that

can effectively utilize longer texts and diverse data sources to enhance overall predictive accuracy. As researchers continue to explore these issues, the potential for improved stock prediction models remains significant.

3.2. ESG Information and Industry-Scenario Extension

As bimodal frameworks gradually matured, researchers began introducing signals with stronger medium- and long-term explanatory value. One significant study combined ESG news sentiment with technical indicators to predict the S&P 500 index, demonstrating that ESG variables not only reflect aspects of corporate governance and sustainability but are also closely associated with investor expectations and market pricing [13]. This integration emphasizes the growing recognition of ESG factors as vital components in financial analysis and highlights their influence on market dynamics.

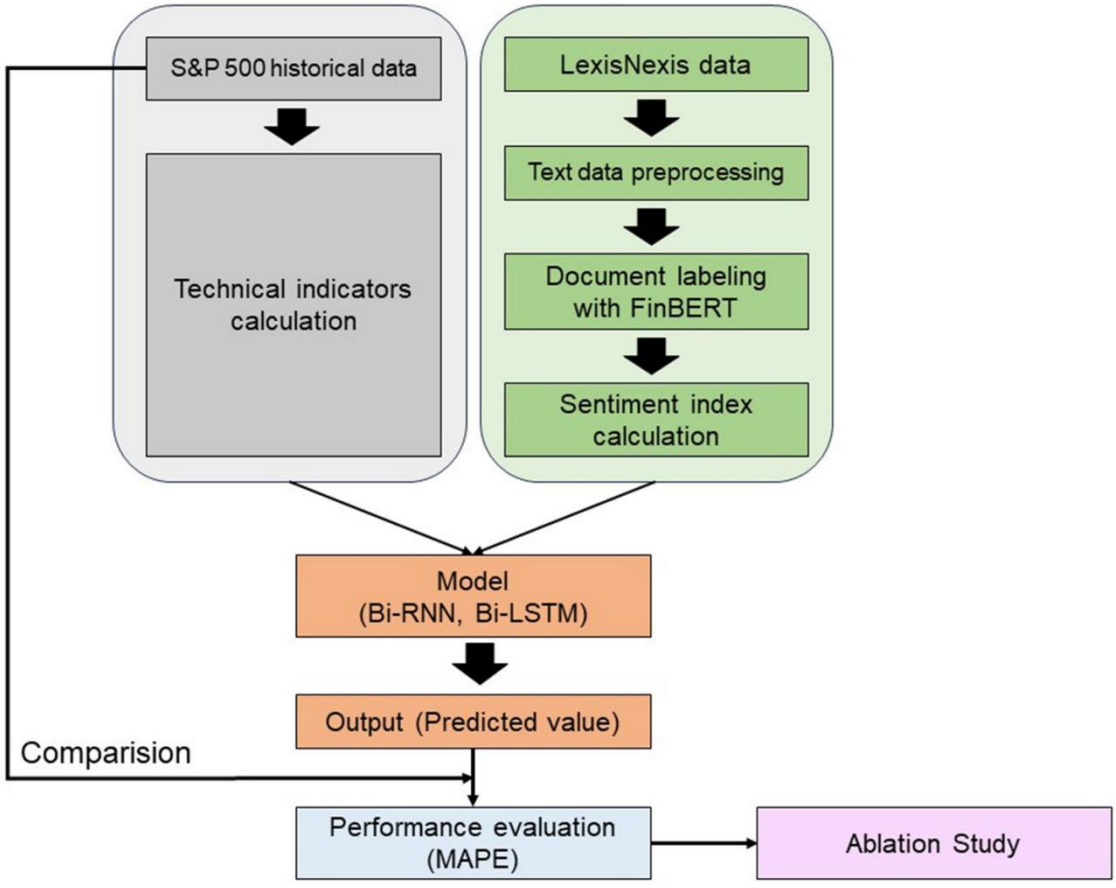

Figure 2. ESG sentiment and technical-indicator prediction pipeline (adapted from [13]).

As shown in Figure 2, historical index data and ESG news text are separated into two distinct processing streams: one dedicated to technical indicator construction and the other focused on sentiment extraction, followed by model-level integration [13]. The content of the image aligns more closely with a pipeline view than with a generic conceptual diagram, effectively

illustrating the flow of information and processes involved in this sophisticated analytical approach. This structured methodology highlights the importance of integrating various data types to enhance predictive modeling in financial markets.

A parallel line of development focuses on industry-specific scenarios. One study concentrated on the Chinese pharmaceutical sector and, in addition to stock time series and news text, introduced COVID-related variables while discussing the influence of exogenous shocks on industry trends [25]. This perspective aligns with event-driven stock prediction research, which emphasizes that market prices are influenced not only by endogenous fluctuations but also by various industry events, policy changes, and public risk factors [5]. By incorporating these considerations, researchers aim to enhance the accuracy of predictions and better understand the dynamics that drive stock performance within specific sectors.

As illustrated in Figure 3, reference [25] presents a multi-stock closing-price chart specifically for firms operating within the pharmaceutical sector. Since the current image depicts a trend plot rather than a comprehensive model framework, the accompanying description is thoughtfully aligned with the observed cross-stock heterogeneity in price trajectories during the pandemic period. This context allows for a deeper understanding of how different pharmaceutical companies responded to the market dynamics, reflecting variations that occurred as a result of shifting economic conditions and public health developments. Overall, this analysis sheds light on the intricate relationships between stock performance and external factors affecting the industry during this critical time.

Figure 3. Closing-price trajectories of pharmaceutical stocks during the pandemic period (adapted from [25]).

From a review perspective, the significance of this stage lies in expanding multimodal stock

prediction from a text-plus-price setting to a framework that also incorporates long-horizon governance signals and exogenous shocks, thereby improving the interpretability of models under real market conditions.

3.3. Graph Relation Modeling and Stable Fusion Mechanisms

As modality types continue to expand, simple feature concatenation becomes increasingly insufficient for effectively representing the complex dependencies across various information sources. The graph attention network proposed offers a versatile tool for stock relation modeling [24]. Later on, a multi-level financial representation framework was introduced that incorporates macro, sector, and stock perspectives [20]. Additionally, another study further enhanced cross-stock relation modeling with a multiscale multimodal dynamic graph convolution network [15]. These advancements reflect the ongoing efforts to develop more sophisticated methods that can capture intricate relationships within multimodal data, ultimately improving the accuracy of stock predictions.

At the fusion level, the MSGCA framework was proposed, which separately encodes indicator sequences, dynamic documents, and relation graphs before integrating them through a two-stage gated cross-attention mechanism [30]. An important characteristic of this design is that relatively stable indicator modalities act as the guiding branch, facilitating a more robust integration process for the noisier and less balanced text and relation modalities. This thoughtful approach enhances the overall effectiveness of the fusion mechanism, allowing for improved performance in stock prediction tasks by leveraging diverse information sources.

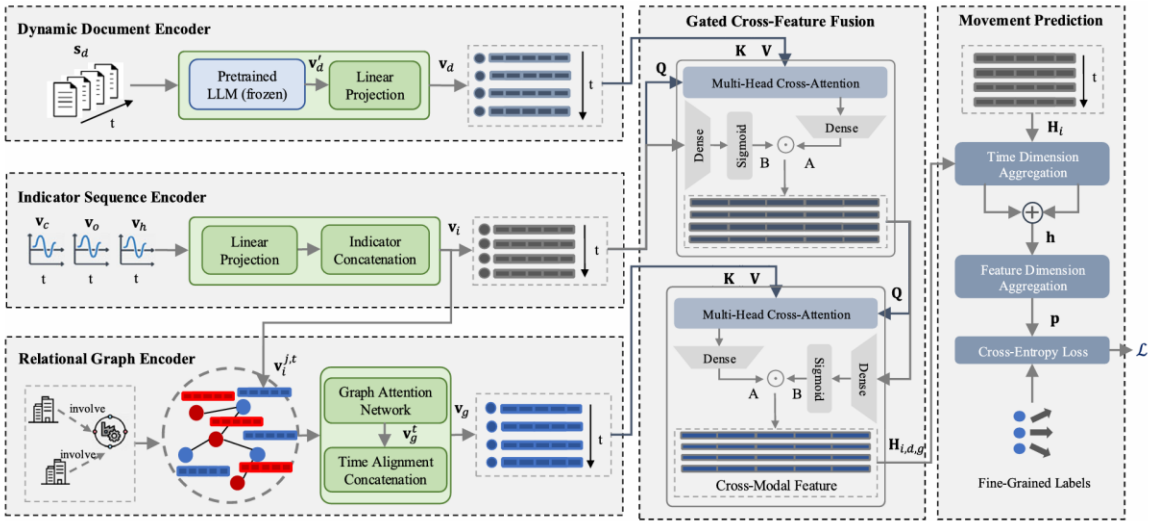

Figure 4. MSGCA framework for stable multimodal fusion (adapted from [30]).

As illustrated in Figure 4, features from indicators, text, and graph relations are progressively

integrated using a sophisticated two-stage gated cross-attention mechanism [30]. This approach ensures that the image content is fully aligned with a robust fusion architecture, rather than merely being represented as an outcome plot. By emphasizing this stable integration, the method effectively enhances the coherence and interpretability of the visual data, allowing for more meaningful insights to be derived from the combined features. Overall, this innovative technique demonstrates significant potential for improved analysis through thoughtful feature integration, paving the way for more accurate stock predictions.

The introduction of stable fusion mechanisms marks a significant shift in research emphasis within the field. Rather than merely questioning whether additional modalities should be incorporated, researchers are increasingly focusing on how to effectively manage challenges such as data sparsity, semantic conflicts, and quality disparities among the various modalities. This evolving perspective highlights a deeper understanding of the complexities involved in integrating multimodal data, paving the way for more robust methodologies that can enhance predictive accuracy and overall performance in diverse applications. As such, addressing these issues has become essential for advancing the effectiveness of multimodal approaches in practical scenarios.

3.4. Long-Text Processing, Patch-Based Modeling, and LLM Expansion

A shared characteristic of financial news, corporate announcements, analytical reports, and social commentary is their increasing length and semantic redundancy. To address this issue, a framework was proposed that combines hierarchical LLM-based text processing with a patch-based Transformer, representing one of the most distinctive recent approaches for handling long texts in multimodal stock prediction [28]. This innovative strategy aims to enhance the efficiency and effectiveness of processing lengthy and complex textual data, ultimately improving the accuracy of predictions by better capturing the relevant information embedded within these extended narratives.

As shown in Figure 5, the system is organized into Hierarchical Summary, Embedding and Alignment, Co-Attention Fusion, and Classification stages [28]. The current image corresponds to a comprehensive modeling framework, aligning well with the surrounding methodological discussion. This structured representation underscores the systematic approach taken in integrating various components for effective long-text handling within multimodal stock prediction, facilitating improved analysis and inference from complex data sources.

This line of work is conceptually continuous with Time-LLM [12], zero-shot time-series forecasting with large language models [9], and related surveys on foundation models for time

series [11]. In all of these studies, large language models are no longer used solely as tools for text understanding; they also participate in time-series modeling through reprogramming or semantic projection.

Meanwhile, tasks such as FinQA [4], DocFinQA [16], and FinTextQA [22] indicate that financial question answering, report understanding, and numerical reasoning over long documents are becoming important directions in financial intelligence. Combined with PatchTST [19], TimesNet [29], and iTransformer [15], these developments are pushing multimodal stock prediction toward a new stage characterized by semantic compression, structural alignment, and stronger explanatory capacity.

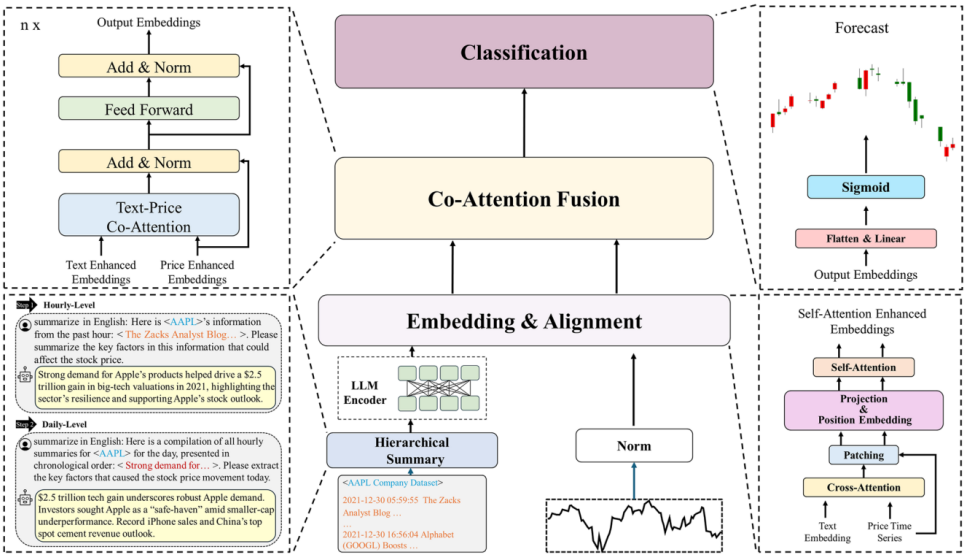

Figure 5. Hierarchical LLM summarization and text-price collaborative modeling framework (adapted from [28]).

4. Major Methodological Categories

Taken together, multimodal stock prediction has developed into a relatively clear and comprehensive technical spectrum. Text-price bimodal fusion serves as the foundational starting point for multimodal modeling, while advancements in ESG factors, industry context, and event enhancement have propelled research toward more complex and realistic market environments. Moreover, the incorporation of graph relations and stable fusion techniques effectively addresses issues related to modality noise and conflicts that may arise during analysis. Additionally, the introduction of Transformers, patch-based modeling approaches, and Large Language Models (LLMs) has significantly expanded the modeling boundaries for both time series data and textual information. This evolution underscores the growing sophistication and adaptability of multimodal methods in capturing the complexities of financial markets.

Table 1. Major methodological categories in multimodal stock prediction.

Method category	Representative studies	Core idea	Main characteristics
Text-price bimodal fusion	[3]; [26]; [21]	Jointly model financial text and historical prices	Establishes the foundation of multimodal prediction
ESG and exogenous enhancement	[13]; [25]; [5]	Introduce ESG, pandemic, policy, and other extended signals	Strengthens interpretability and scenario adaptation
Graph relations and stable fusion	[30]; [24]; [20]; [15]	Use graph structures and gating mechanisms to integrate multi-source information	Addresses modality conflict and relational dependence
Transformer and patch-based temporal modeling	[23]; [19]; [29]; [15]	Replace conventional RNNs with efficient long-sequence modeling	Improves temporal representation capacity
LLMs and long-document processing	[28]; [12]; [9]; [27]; [4]; [16]; [22]	Enhance semantic modeling through summarization, reprogramming, and reasoning	Extends to long-document understanding and richer explanations

5. Key Challenges

5.1. Insufficient Label Granularity and Task Definition

Although multimodal stock prediction studies have continued to expand in both the number of modalities and model complexity, most tasks are still formulated as binary up-down classification or coarse up-flat-down classification. Such label definitions are convenient for standardized evaluation, yet they remain insufficient for representing finer market states such as flat movement, accelerated rise, accelerated fall, and near-reversal conditions.

5.2. Long-Text Redundancy and Insufficient Explicit Event Modeling

Financial news and corporate announcements often contain substantial background information, repetitive narrative, and content that is not directly price sensitive. Although hierarchical LLM summarization [28] and long-document financial tasks such as FinQA [4], DocFinQA [16], and FinTextQA [22] provide new processing routes, comparatively few stock-prediction studies explicitly model event type, impact strength, transmission window, and market feedback path in an integrated manner.

5.3. Modality Conflict, Missing Data, and Fusion Stability

Different modalities are not always consistent with one another. Prices may decline while textual signals remain positive; a highly salient event may receive little immediate price

response; and some stocks may suffer from sparse news coverage or weak graph connectivity. Consequently, multimodal fusion is not merely a feature-concatenation problem but also a modality-reliability problem. Works such as MSGCA [30] already recognize this issue, yet stable fusion remains a key challenge that has not been fully resolved.

5.4. Interpretability and Cross-Market Transfer

Financial prediction tasks inherently require strong interpretability. Investors and researchers are concerned not only with whether a prediction is accurate but also with why a model reaches a particular conclusion. However, many current approaches still lack clear explanations of the textual evidence, event drivers, and relation-propagation paths underlying their predictions. At the same time, transferability across markets, sectors, and time periods remains under-discussed.

6. Summary of Research Trends

The existing literature indicates that the main trajectory of multimodal financial stock prediction is becoming increasingly refined.

First, task definitions are moving toward finer-grained state modeling, with growing attention to fine-grained state prediction, reversal identification, and joint risk-return modeling.

Second, explicit event modeling is becoming more closely integrated with long-document understanding, making the interaction among news, announcements, financial reports, and event graphs an important extension of recent research.

Third, stable fusion and interpretability have become more central evaluation dimensions. Reporting accuracy alone is no longer sufficient to reflect the practical value of a model; the ability to explain predictive evidence, recognize modality conflict, and maintain temporal stability is increasingly treated as a marker of research quality.

Fourth, financial large language models and foundation models are significantly propelling the field toward more unified modeling approaches and a deeper understanding of long documents. Models such as FinGPT [27] and Time-LLM [12], along with related studies like [9] and [11], are reshaping the joint modeling of textual information and time series data, fostering a more cohesive analytical framework. Meanwhile, initiatives like FinQA [4], DocFinQA [16], and FinTextQA [22] are accelerating the integration of stock prediction techniques with comprehensive financial document understanding. This convergence not only enhances predictive capabilities but also enriches the contextual insights that can be drawn from complex financial narratives, highlighting the transformative potential of these advanced models in the field of finance.

7. Conclusion

Centered on five recent representative papers and supported by related studies on financial text representation learning, graph relation modeling, patch-based temporal modeling, and financial large language models, this paper has reviewed the research progress of multimodal stock prediction. From an overall perspective, the field has evolved from text-price bimodal fusion to joint modeling over ESG, exogenous events, graph relations, and long-text semantic extraction.

At the same time, insufficient label granularity, long-text redundancy, limited explicit event modeling, unstable multimodal fusion, and insufficient interpretability remain core problems in current multimodal financial prediction. The overall direction visible in the literature shows that finer task definitions, stronger event and relation modeling, and more robust cross-modal fusion mechanisms are becoming the key themes in the continuing evolution of this field.

References

[1] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv, arXiv*:1908.10063.

[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, *2(1), 1-8*.

[3] Chen, P., Boukouvalas, Z., & Corizzo, R. (2024). A deep fusion model for stock market prediction with news headlines and time series data. *Neural Computing and Applications*, 36, 21229-21271.

[4] Chen, Z., Chen, W., Smiley, C., et al. (2021). FinQA: A dataset of numerical reasoning over financial data. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3697-3711).

[5] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2327-2333).

[6] Dong, Z., Fan, X., & Peng, Z. (2024). FNSPID: A comprehensive financial news dataset in time series. *arXiv, arXiv*:2402.06698.

[7] Feng, F., Chen, H., He, X., Ding, J., Sun, M., & Chua, T.-S. (2019). Enhancing stock movement prediction with adversarial training. *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 5843-5849).

[8] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654-669.

[9] Gruver, N., Finzi, M., Qehaja, M., & Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. *In Advances in Neural Information Processing Systems*.

[10] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv, arXiv*:2312.00752.

[11] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., & Wen, Q. (2023). Time series forecasting with pretrained foundation models: A survey. *arXiv, arXiv*:2312.17119.

[12] Hu, Z., Zhao, Y., & Khushi, M. (2023). Time-LLM: Time series forecasting by reprogramming large language models. *arXiv, arXiv*:2310.01728.

[13] Lee, H., Kim, J. H., & Jung, H. S. (2024). Deep-learning-based stock market prediction incorporating ESG sentiment and technical indicators. *Scientific Reports*, 14, Article 10262.

[14] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted transformers are effective for time series forecasting. *In Proceedings of the International Conference on Learning Representations (ICLR).*

[15] Liu, Z., Chen, C., Wang, B., et al. (2024). Melody-GCN: A multiscale multimodal dynamic graph convolution network for stock forecasting. *Pattern Recognition*, 61(1).

[16] Lu, Y., Xing, H., & Zhou, C. (2024). DocFinQA: A long-context financial reasoning dataset. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[17] Maia, M., Handschuh, S., Freitas, A., et al. (2018). WWW'18 open challenge: Financial opinion mining and question answering. *In Companion Proceedings of the Web Conference 2018*.

[18] Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782-796.

[19] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv, arXiv*:2211.14730.

[20] Qin, W., Song, Y., Chen, W., et al. (2021). Coupling macro-sector-micro financial indicators for learning stock representations with less uncertainty. *In Proceedings of the 35th AAAI Conference on Artificial Intelligence* (pp. 9122-9129).

[21] Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8415-8426).

[22] Shen, T., Zhang, S., Li, Y., et al. (2024). FinTextQA: A dataset for long-form financial question answering. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[23] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems (Vol. 30)*.

[24] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv, arXiv*:1710.10903.

[25] Wang, H., Xie, Z., Chiu, D. K. W., & Ho, K. K. W. (2025). Multimodal market information fusion for stock price trend prediction in the pharmaceutical sector. *Applied Intelligence*, 55, Article 77.

[26] Xu, Y., & Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1970-1979).

[27] Yang, Y., Xie, Y., Wang, C., et al. (2023). FinGPT: Open-source financial large language models. *arXiv, arXiv*:2306.06031.

[28] Zhang, Y., Dong, Z., & Xu, W. (2026). Integrative stock price trend prediction via hierarchical LLM text processing and patch-based transformer with co-attention. *Expert Systems with Applications*, 302, Article 130441.

[29] Zhou, T., Zhang, P., & Yan, Z. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[30] Zong, C., Wan, J., Cascone, L., & Zhou, H. (2025). Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *Complex & Intelligent Systems*, 11, Article 396.

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Flexible blank removal robot for ceramic production line

Zihou Wu, Zixiao Wang*, Yulin Zhang, Liqian Cai, Jialin Yang, Zhicheng Liang, Ruosi Li, Haining Huang

Quanzhou Vocatonal and Technical College

Received: May 25, 2026

Revised: May 26, 2026

Accepted: May 27, 20265

Published online: May 29, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 6 (June 2026)

* Corresponding Author: Zixiao Wang (754198970@qq.com)

Abstract. The manual handling of wet ceramic greenware after slip casting results in low efficiency, high breakage rates, and harsh labor conditions. To address these issues, this paper presents the design and simulation of a flexible greenware-picking robot tailored for slip casting production lines. An articulated rotary coordinate structure with six degrees of freedom is adopted. Key components including the forearm, wrist, and rotary base are designed and calculated. A vacuum suction cup matrix is selected as the end effector, and a harmonic reducer is integrated for precise torque transmission. A digital prototype is constructed using SolidWorks, followed by kinematic and force simulations. Simulation results verify that the robot achieves a stable grasping force of 350 N distributed across four suction cups, a rated payload of 5 kg, and positioning accuracy of ±3–5 mm. The robot operates within a workspace of 2500 mm × 1200 mm × 1100 mm and meets the cycle time requirements of typical slip casting lines. This work provides a verified paradigm for robotic automation in the ceramic industry.

Keywords: *Slip Casting; Robotic Grasping; Vacuum End Effector; Harmonic Drive; Structural Simulation*

1. Introduction

During the traditional ceramic manufacturing process, the wet greenware formed via slip casting must be demolded and transferred carefully to subsequent processing stations. Currently, the vast majority of ceramic enterprises heavily depend on manual labor to perform these repetitive picking and placement operations. This specific production stage is notoriously characterized by an exceptionally humid operating environment and high labor intensity. Furthermore, due to the brittle and soft nature of wet greenware, manual handling frequently introduces uneven contact forces or accidental structural collisions, leading to high product

breakage rates and significantly undermining the overall yield of qualified products. With the continuous inflation of human resource costs and the strategic national push toward intelligent manufacturing systems, the implementation of comprehensive process automation in the greenware-picking stage has emerged as an urgent industrial demand.

Conventional rigid industrial robots fail to adequately adapt to the significant geometrical variability and wet, soft material attributes of freshly cast ceramic greenware, which inherently leads to severe structural damage or grasping failures. Conversely, a specialized flexible greenware-picking robot equipped with inherent compliance and self-adaptive adjustments can smoothly and accurately handle distinct topologies of wet greenware without inflicting surface defects. This research is directly derived from the practical automation upgrade needs of the ceramic sanitary ware sector. It aims to develop a dedicated flexible robotic system to resolve the low efficiency and high rejection rates associated with manual operations, facilitating a thorough intelligent reconstruction of the manufacturing line.

1.1. Research Objectives and Significance

This investigation focuses explicitly on the slip casting ceramic production environment to design, calculate, and simulate a flexible greenware-picking robot. The core objectives include:

- Elucidating the physical material properties of wet ceramic greenware and mapping the corresponding pick-and-place process prerequisites. By analyzing the structural dimensions, mass limits, and cycle rhythms, the necessary payload, positioning precision, movement velocity, and flexible adaptation parameters are quantified to establish rigorous inputs for mechanical engineering.
- Developing the comprehensive mechanical configuration of the flexible picking robot. Based on a multi-joint articulated configuration, the integrated structure comprising the base, major arm, forearm, wrist, and customized end effector is engineered with an emphasis on incorporating structural compliance to cushion interactive forces.
- Performing parametric optimization and mechanical verification of critical load-bearing assemblies. Drive components such as motors and reducers are mathematically selected based on dynamic requirements, and structural elements are verified using static mechanics to guarantee absolute structural integrity.
- Constructing a high-fidelity 3D digital prototype and executing multi-body kinematic simulations to analyze operational trajectories and workspace profiles.

The significance of this study spans both academic and practical dimensions. Theoretically,

robotic handling specialized for highly fragile and soft workpieces like wet ceramic greenware remains vastly under-researched. This paper integrates compliant control theories with empirical ceramic process parameters, offering novel design methodologies for non-standard, damage-prone industrial workpieces. Practically, the engineered robot can be directly deployed on sanitary ware lines, liberating workers from hazardous environments, securing stable cycle times, minimizing material waste, and advancing the green manufacturing paradigm.

1.2. Literature Review of Related Works

Domestic researchers have achieved valuable progress in applying robotic technology to post-processing ceramic stages. For instance, Li et al. analyzed the macro trends of industrial robots driving efficiency enhancements and cost minimization across the ceramic sector. Regarding specific process control, Li and Zhou investigated time-scaled industrial robotic control strategies to optimize trajectories during ceramic glazing operations. Liu explored the application of single-chip microcomputers in sanitary ceramic glazing robots to enforce quality consistency, while He designed an automated glazing workstation to substitute manual labor. In terms of automated polishing, Zhang developed an automatic path-planning algorithm based on 3D machine vision and point cloud extraction for ceramic greenware grinding, and Liu further optimized these vision-guided polishing trajectories. Chen conducted thick glazing simulations and trajectory optimization for sanitary ware to ensure coating uniformity. Advanced force control systems have also been studied; Yang implemented a hybrid force/position control mechanism for ceramic grinding, and Hong engineered a constant-force actuator tailored for robotic polishing of sanitary greenware. Huang and Tian further improved closed-loop glazing and simplified programming methods for teach-pendant systems. Nevertheless, these studies are primarily constrained to mid-stream and downstream processes such as glazing and polishing. Dedicated robotic systems for the upstream slip casting and wet greenware handling stages are exceedingly rare, with most existing solutions merely modifying general-purpose manipulators without fundamental structural innovation for high compliance. International research in this domain exhibits distinct focus areas, primarily emphasizing the intersection of advanced computational mechanics and intelligent material design. Zhenhao et al. successfully leveraged machine learning architectures to optimize the development of Al2O3-SiO2 porous ceramics based on small-sample datasets. Similarly, Ghaffari et al. formulated verification workflows for complex ceramic machine learning interatomic potentials to improve computational molecular simulations. On the manufacturing front, Zhu et al. conducted extensive numerical simulations to investigate crack initiation and suppression

mechanisms during robot-assisted abrasive belt grinding of zirconia ceramics, establishing precise chip thickness models. Collectively, international literature focuses on material intelligence and fine-grained micro-machining physics. Specialized robotic system design prioritizing macro structural flexibility and reliability for high-moisture greenware handling remains limited, although high-end sensors and precision components provide vital references for this research. Furthermore, to validate the structural feasibility and kinematic reliability of the proposed mechanical architecture, this study constructs a high-fidelity digital prototype using SolidWorks. Comprehensive kinematic and dynamic simulations are executed to evaluate the performance of key transmission components, particularly the selected harmonic reducers. The simulation results aim to verify that the robotic system can achieve a stringent positioning accuracy of ±3–5 mm and a highly efficient operation cycle of 4.2 seconds, fully satisfying the high-throughput demands of modern automated production lines. The remainder of this paper is organized as follows: Section 2 delineates the kinematic modeling and the overall structural architecture of the 6-DOF manipulator. Section 3 details the mechanical design of critical components, with a specific focus on the harmonic reducer selection and the mathematical modeling of the flexible vacuum suction end-effector. Section 4 presents the SolidWorks-based simulation framework, analyzing the kinematic trajectories and dynamic stress distributions. Finally, Section 5 draws the conclusions and outlines potential avenues for future research.

2. Methodology

2.1. Fundamental Design Requirements

To ensure the successful deployment of the automated manipulation system within the ceramic manufacturing facility, the structural layout and control architecture of the flexible greenware-picking robot must strictly accommodate the harsh production conditions and complex wet material traits. The engineering constraints can be categorized into six core dimensions:

2.1.1. Material Rheology and Compliance Requirements

The newly formed wet ceramic greenware exhibits unique rheological properties, characterized by low structural rigidity, high moisture content, and extremely high sensitivity to external mechanical stress. Unlike dried or fired ceramic bodies, unfired clay possesses a low yield strength, making it highly susceptible to micro-cracking, local surface indentation, or catastrophic structural collapse under non-uniform gripping forces. Therefore, the robotic manipulator and its end-effector must exhibit outstanding structural compliance. The picking

mechanism must adapt to the minor geometric irregularities of the greenware surface, ensuring that the contact forces are evenly distributed and safely below the material's critical deformation threshold.

2.1.2. Dynamic Stiffness and Vibration Suppression

Concurrently, the mechanical framework of the robot must maintain a high level of structural stiffness and dynamic rigidity. During high-speed transfer operations between the casting mold and the subsequent processing line, the robotic links are subjected to intense inertial forces and alternating dynamic loads. High structural stiffness is paramount to suppressing excessive mechanical vibrations, preventing structural resonance, and minimizing settling times at the end of a trajectory. Any uncontrolled oscillation of the robot arm would be transmitted directly to the fragile greenware, leading to severe inertial shear stresses that could damage the internal microstructure of the wet clay.

2.1.3. Kinematic Dexterity and Workspace Optimization

The spatial layout of traditional slip-casting workshops is highly congested, featuring multi-layered, densely arranged plaster molds and complex auxiliary pipe networks. Consequently, the robot must possess sufficient Degrees of Freedom (DOFs)—specifically an articulated Six-Degree-of-Freedom (6-DOF) rotary coordinate structure—to smoothly navigate complex mold layouts and multi-angle extraction paths. The kinematic configuration must provide an optimized operational envelope while avoiding kinematic singularities and joint limit collisions. This dexterity allows the robot to approach the molds from optimal entry vectors, execute smooth extraction trajectories, and avoid any physical interference with neighboring manufacturing equipment.

2.1.4. Positioning Accuracy and Trajectory Correction

In high-throughput automated ceramic production, positioning accuracy and repetitive precision are critical engineering metrics that can significantly impact operational efficiency. Even minor misalignments during the picking or placing phase can lead to collisions with the mold walls or off-center handling, both of which may result in immediate structural damage to the greenware. To effectively mitigate this risk, the robotic control system must incorporate high-resolution feedback loops that are capable of dynamically correcting trajectories in real-time. This sophisticated system must also compensate for structural deflections caused by gravity, link elasticity, and joint clearance, thereby ensuring consistent, high-precision path tracking even under varying payload conditions. Such advancements are essential for optimizing production quality and minimizing waste in the manufacturing process.

2.1.5. Non-Concentrated Stress Distribution of the End-Effector

The end-effector serves as the critical physical interface between the robotic system and the delicate ceramic workpiece. To lift and support the rated load without causing localized structural failure, the end-effector must deliver uniform vacuum suction. Traditional concentrated or point-source gripping configurations must be strictly avoided. Instead, a distributed matrix architecture must be utilized to diffuse the holding forces across multiple independent adsorption nodes. By optimizing the spatial distribution of these suction points, the negative pressure can be controlled to provide a robust total lifting force while ensuring that the localized stress concentrations on the wet greenware surface remain practically negligible.

2.1.6. Environmental Resistance and Maintainability

Finally, the operational environment of slip-casting shops presents severe challenges to the long-term reliability of industrial machinery. The atmosphere is characterized by sustained high humidity, ambient temperature fluctuations, and highly corrosive agents derived from the chemical composition of the ceramic slurry and mold-release factors. Therefore, all critical sub-assemblies—including the harmonic reducers, stepping motors, encoder feedback units, and structural joints—must undergo rigorous moisture-proofing, dust-proofing (IP65 rating or higher), and anti-corrosion surface treatments. Furthermore, the overall mechanical design must emphasize modularity to facilitate straightforward maintenance, quick component replacement, and minimized industrial downtime.

2.2. Mechanical Structure Evaluation and Selection

To realize the mandatory 6-DOF motion capability, three distinct structural configurations were thoroughly evaluated:

- Gantry/Framework Structure: This configuration relies on large-scale linear rails to move the entire arm block. While it offers superior structural rigidity and heavy payload capabilities, its vast spatial footprint and low dexterity render it highly incompatible with the congested layouts of modern slip casting lines.
- Spherical/Polar Coordinate Structure: This architecture mimics spherical motion via rotation, pitch, and extension. Although compact, its complex mechanical linkages and transmission paths are highly prone to rapid wear and precision degradation when continuously exposed to abrasive ceramic dust and moisture.
- Articulated Rotary Coordinate Structure: Modeling human arm kinematics, this layout utilizes a series of revolute joints supported by a robust column. It delivers exceptional

dexterity, an expansive workspace envelope, and high reliability. It can smoothly adjust orientations to match irregular mold configurations.

2.3. Rated Payload and Technical Parameters

The rated payload directly governs the operational stability of the system. Based on standard production specifications, the mass of an individual wet ceramic greenware piece ranges from 1.5 kg to 3.0 kg. Accounting for the self-weight of the vacuum end effector, integrated force sensors, and compliant buffering links, the total static load at the terminal interface is approximately 4.0 kg. Applying an industrial safety coefficient of 1.25 to 1.5 to counteract dynamic inertial forces during rapid acceleration/deceleration, the rated payload is established at 5.0 kg. The key performance parameters are compiled in Table 1.

Table 1. An example table.

Performance Indicator	Primary Engineering Parameter
Number of Degrees of Freedom (DOF)	6
Rated Payload Capacity	5.0 kg
Operational Workspace Volume	2500 mm × 1200 mm × 1100 mm
Static Positioning Accuracy	±3 to 5 mm

2.4. Structural Composition

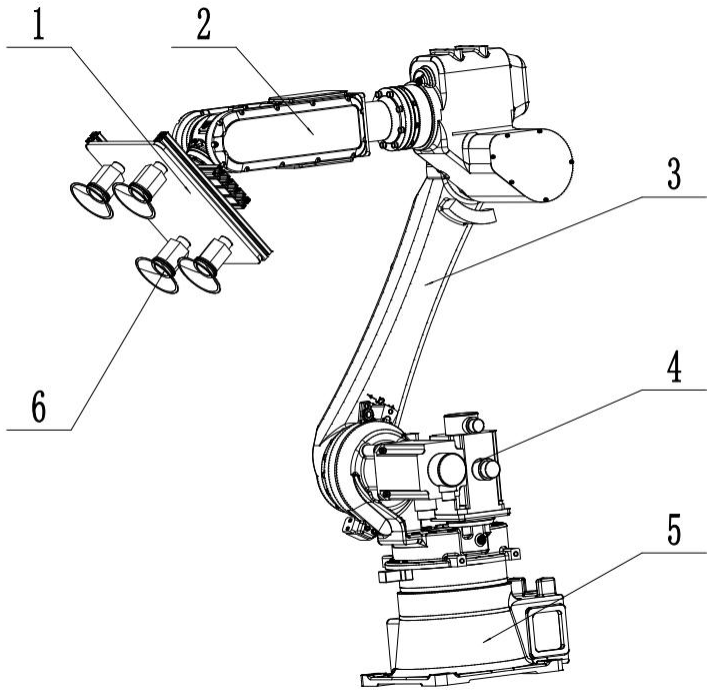

Figure 1. Structural schematic of the flexible greenware-picking robot for a slip-casting ceramic production line.

Figure 1 illustrates the structural schematic of the flexible greenware - picking robot for a

slip - casting ceramic production line, where component 1 is the end - effector assembly, component 2 is the forearm mechanism, component 3 is the upper - arm mechanism, component 4 is the drive motor and harmonic reducer, component 5 is the base platform, and component 6 is the vacuum suction cup.

The robot is structurally composed of a heavy-duty base, a major arm, a forearm, a high-precision transmission assembly, stepping motors, harmonic reducers, and a vacuum-based end effector. The base acts as the foundational support, fabricated from HT250 gray cast iron to provide high mass and vibration damping. The major arm links the base to the forearm and provides primary elevation; it is engineered with a hollow rectangular cross-section using 6061 aluminum alloy to balance lightweighting with high bending resistance. The forearm controls detailed spatial positioning, fabricated from high-grade Q235 structural steel to minimize rotational inertia while maintaining rigidity. Dynamic actuation is delivered via specialized stepping motors coupled with high-reduction harmonic drives (composed of a wave generator, flex spline, and circular spline) using 40Cr alloy steel with induction-hardened teeth. The end effector consists of a multi-point nitrile rubber vacuum suction cup matrix connected to an optimized vacuum generator, ensuring damage-free grip on moist surfaces.

2.5. Mechanical Characterization of the Forearm

The forearm is the core kinematic component of the flexible greenware-picking robot in the slip-casting ceramic production line, responsible for driving the end effector to execute the grasping and placement operations of wet greenware. To ensure reliable performance, the mechanical design of the forearm must satisfy the following critical requirements:

2.5.1. High Structural Rigidity and Kinematic Stability

Ceramic wet greenware is inherently soft and brittle; therefore, any microscopic vibration or structural deformation during the picking and placing process may induce severe damage to the greenware body. The forearm material should be selected from structural materials characterized by high strength and low density, such as Q235 steel, to minimize the operational rotational inertia while guaranteeing sufficient payload capacity. The forearm adopts a cantilever or linkage configuration, optimized through rigorous cross-sectional geometries and reinforcement ribs, to ensure excellent morphological stability during continuous operations and prevent any degradation in picking accuracy caused by unexpected structural deformation.

2.5.2. Flexible Multi-Degree-of-Freedom (Multi-DOF) Motion

Given the variations in the initial position and spatial orientation of the wet greenware within

the molds, the forearm must be capable of multi-directional position and angular adjustments. The forearm connects the major arm and the wrist mechanism via rolling joints, constituting a multi-axis kinematic chain. This configuration enables the vacuum end effector to reach any designated position within the workspace and seamlessly adapt to diverse mold layouts and picking postures.

2.5.3. High Precision and Rapid Response of the Transmission System

A high-performance servo motor coupled with a precise harmonic reducer is implemented for power transmission to enhance torque density and positioning accuracy. A closed-loop control system monitors the forearm position in real time and dynamically updates the motion trajectory, ensuring accurate and highly repeatable picking actions. Furthermore, the layout of the transmission components must be compact to minimize the spatial footprint, thereby facilitating installation and maintenance within the constrained production line environment.

2.5.4. Balanced Lightweighting and High Robustness

On the premise of satisfying the payload and rigidity boundaries, topology optimization is conducted to effectively reduce the net mass of the arm link. This reduction not only lowers the loads on the joint motors but also enhances motion response speed and improves overall energy efficiency. Furthermore, specialized anti-moisture and anti-corrosion surface treatments are applied to the forearm, ensuring its durability and resilience in withstanding the aggressive and humid production environment that is characteristic of the ceramic slip-casting process. These measures collectively contribute to the operational reliability and longevity of the forearm mechanism in demanding conditions.

Figure 2 illustrates the structural schematic diagram of the proposed forearm mechanism, which primarily comprises the driving motor, the transmission assembly, and the forearm main body.

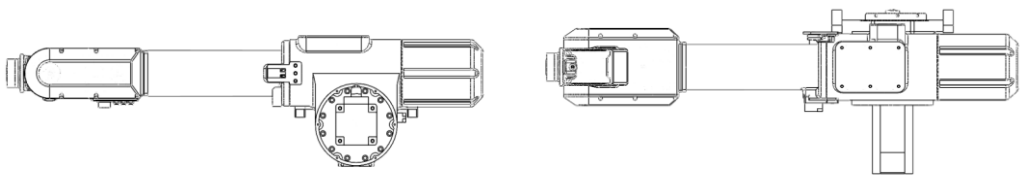

Figure 2. Forearm Mechanism.

The forearm mechanism primarily consists of the driving motor, the transmission assembly, and the main body of the forearm. To ensure sufficient rigidity and durability, structural Q235 steel is selected as the primary material for the arm body. Based on the physical properties

obtained from the Three-Dimensional (3D) digital prototype modeling, it has been determined that the total mass of the integrated forearm assembly is approximately 23.9 kg, as illustrated in Figure 3. This weight consideration is crucial for optimizing performance and ensuring the overall functionality of the forearm mechanism in various applications.

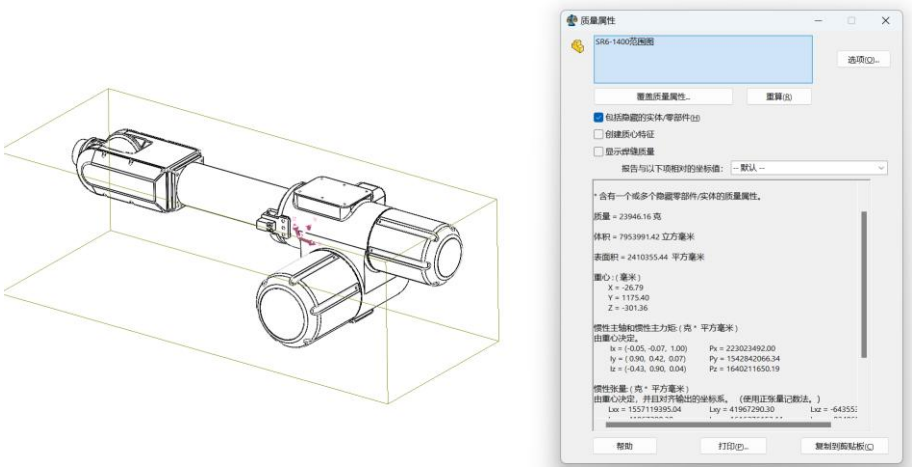

Figure 3. Forearm Mass.

According to the system design requirements, the mass moments of inertia of the dual arms and the wrist assembly ($J_{G1} $, $J_{G2} $, and $J_{G3} $) are established based on their respective center-of-gravity axes and the palletizing axis. To evaluate the total rotational inertia of the entire arm mechanism about the first joint axis, the parallel axis theorem is applied. The total moment of inertia is mathematically formulated by incorporating the distance between the center-of-gravity axis and the first joint axis:

$$J_1 = J_{G1} + m_1 {l_1}^2 + J_{G2} + m_2 {l_2}^2 + J_{G3} + m_3 l_3^2$$

$$J_1 = m_1 {l_1}^2 + m_2 {l_2}^2 + m_3 l_3^2 = 10 \times 0.185^2 + 5 \times 0.8^2 + 12 \times 1.5^2 = 30.54 kg.㎡$$

The moment of inertia about the second joint axis (the main stacking pivot) is:

$$J_2 = m_2 {l_4}^2 + m_3 {l_5}^2 = 5 \times 0.4^2 + 12 \times 0.8^2 = 8.48 kg \cdot m^2$$

Therefore, the initial startup torque required by the mechanism at the onset of the palletizing operation is formulated as follows:

$$T = J \times \omega$$

If the angular velocity of the main axis needs to be adjusted from its current velocity to the target angular velocity within a duration of $\Delta t = 1s$, then:

$$T_1 = J_1 \times w = J_1 \times \frac{w - w_0}{\Delta t} = 30.54 \times \frac{1.22\pi}{1} \approx 116.9 N \cdot m$$

Therefore, this paper determined the magnitude of the moment of inertia by calculating the

values for both arms and the wrist around the second joint axis, which serves as the palletizing axis. Furthermore, if the moment of inertia around the center of mass axis for each individual section of the robotic arm, as well as the friction torque, are comprehensively taken into account, the startup torque at the onset of the palletizing operation can be assumed to be "120N·m".

The power output of the electric motor can be estimated based on the mathematical formula provided below:

$$P_m \approx (1.5 \sim 2.5) \frac{M_{LP} \Omega_{LP}}{\eta}$$

Where:

P_m represents the power output of the electric motor (W);

M_{LP} denotes the magnitude of the load torque $(N \cdot m)$;

Ω_{LP} signifies the operational speed of the load (rad/s);

η indicates the transmission efficiency.

Based on the aforementioned calculation data, the standard power of the electric motor can be selected to ensure that the motor power is maximized. The rated power PR must satisfy the mathematical formula listed below:

$$P_r \geq P_m$$

Therefore, a series-wound DC motor is selected as the actuation source, and the specific motor model is designated as QZD-08.

Table 2. Technical specifications of the QZD-08 drive motor.

Parameter Description	Standard Value
Rated Power Capacity (PR)	800 W
Rated Operational Voltage	24 V
Rated Current Draw	46.2 A
Rated Spindle Speed	1750 r/min
Excitation Mode / Insulation Class	Series Excitation / Class B
Standard Duty Cycle Rating	60 min

2.6. Motor and Reducer Selection for Forearm Joints

Given that the forearm mechanism of the flexible greenware-picking robot in the ceramic slip-casting production line possesses Two Degrees of Freedom (2-DOF), it primarily operates within a planar palletizing motion workspace. Owing to the negligible static friction between the robot frame and the bearings, coupled with the well-conditioned surface roughness, the static torque required for palletizing remains relatively low.

However, based on analytical mechanics, when the arm configuration extends completely into a straight line, the mass moment of inertia reaches its maximum value, causing a concomitant and sharp increase in the torque demanded from the stepping motor. Consequently, for operational scenarios with demanding requirements for heavy payloads and high velocities, separating the entire joint into a multi-link or two-stage articulation mechanism presents a viable design alternative.

Herein, all relevant mass moments of inertia *(J_{G1},J_{G2}, and J_{G3})* are rigorously defined relative to the center-of-gravity axes and the palletizing axis. Once the perpendicular distance between the center-of-gravity axis and the first joint axis is geometrically determined, the equivalent rotational inertia of the dual arms about the first joint axis can be explicitly calculated using the parallel axis theorem as follows:

$$J_1 = J_{G1} + M_1{L_1}^2 + J_{G2} + M_2{L_2}^2 + J_{G3} + M_3L_3^2$$

Substituting the structural design data into the equation yields:

$$J_1 = M_1{L_1}^2 + M_2{L_2}^2 + M_3L_3^2 = 4 \times 0.143^2 + 1 \times 0.445^2 + 4 \times 0.542^2 = 1.46kg \cdot m^2$$

$$T = J \cdot dt/d\omega = 2.71 \times 0.215 \times \pi/180 \approx 3.54\ N \cdot m$$

Therefore, this paper determines the precise magnitude of the mass moment of inertia by calculating the values for both arms and the wrist assembly, utilizing the second joint axis as the designated palletizing axis:

$$M_2 = 1kg, L_4 = 97mm, M_3 = 4kg, L_5 = 194mm$$

$$J_2 = M_2{L_4}^2 + M_3{L_5}^2 = 1 \times 0.097^2 + 4 \times 0.194^2 = 0.16kg \cdot m^2$$

Given that the angular velocity of the robotic forearm shifts from the initial state to *15°/s*, the required time duration for this transition is approximately 0.2s.

$$T_2 = J_2 \times \dot{\omega_2} = 0.16 \times \frac{\pi/12}{0.2} = 0.21N \cdot m$$

To guarantee operational reliability, a startup torque of 10N·m at the beginning of the palletizing process is taken into account. By incorporating a safety factor of 2.

$$T01 = 2T = 2X10 = 20N.m$$

Based on the aforementioned calculation data, the XB3-50-100 model is selected as the harmonic reducer for the robotic system. By consulting the specification catalog of the harmonic reducer, the standard output torque is determined to be *$20N{\cdot}m$*, and the transmission gear ratio is designated as *i_1=100.*

According to the parameter sheet, the transmission efficiency of the harmonic reducer is

specified as η_1=*0.9*. Consequently, the minimum required driving torque T_m that the stepping motor must output can be calculated as follows:

$$T_{out1} = \frac{T_{01}}{i \cdot \eta} = \frac{20}{100 \times 0.9} = 0.22N \cdot m$$

Consequently, a variable reluctance stepping motor of the BF series is chosen as the drive source, with the specific motor model designated as 55BF003.

By consulting the technical parameter specification sheet of this stepping motor, its rated holding torque is determined to be 0.686N·m, and the step angle is specified as 1.5˚.

Since the maximum holding torque *(0.686N·m)* is significantly greater than the minimum required driving torque *(0.22N·m)* calculated previously, this selected motor possesses an adequate torque margin, thereby fully satisfying the dynamic requirements during the startup phase of the palletizing operation."

2.7. Selection of Transmission Mechanism Types

Given that this system not only requires a high transmission ratio but must also possess sufficient load-bearing capacity, this paper, after thoroughly studying relevant literature, decides to utilize a reducer to meet this requirement. In this design, this paper selects the wave generator as the driving component, while designating the flexible spline as the driven component. The primary reason for this choice is that the pressure angles of the gears within the reducer are uniformly set to the standard value of *20°*."

2.8. Parameter Selection of Reducer

Determination of Tooth Numbers

According to the relevant design literature of harmonic reducers, the tooth number of the flexible spline satisfies:

$$Z_r = u \times i = 2 \times 100 = 200$$

According to the relevant design literature of harmonic reducers, the tooth number of the circular spline satisfies:

$$Z_g = u + Z_r = 2 + 200 = 202$$

Given that the module is $m = 0.5mm$, then:

Through calculation, the pitch diameter of the flexible spline can be obtained as:

$$d_r = mZ_r = 0.5 \times 200 = 100mm$$

Through calculation, the pitch diameter of the circular spline can be obtained as:

$$\delta_1 = (75 + {}^{Z_r}/_4)d_r \times 10^{-4} = (75 + {}^{200}/_4) \times 100 \times 10^{-4} = 1.25mm$$

Therefore, the calculated value of $\Delta 1$ can be increased by 20 times. Consequently, under heavy-load conditions, higher strength is required for the reducer. To enhance the stiffness of the gearbox, the diameter of the plunger needs to be increased.

$$\delta_1 = 1.25 \times 1.20 = 1.5mm$$

Through calculation, the body wall thickness of the flexible spline can be obtained as:

$$\delta = 0.7\delta_1 = 0.7 \times 1.5 = 1.05mm$$

To improve the stiffness of the flexible spline, the value can be appropriately increased, yielding $\delta = 1.2mm$.

Through calculation, the tooth width is obtained as:

$$B = 0.15d_r = 0.15 \times 100 = 15mm$$

Through calculation, the length of the hub flange is obtained as:

$$C = (0.2\sim0.3)B = (0.2\sim0.3) \times 15 = 3\sim4.5mm，\ C = 4mm$$

Through calculation, the body length of the flexible spline is obtained as:

$$L = (0.8\sim1.2)d_r = (0.8\sim1.2) \times 100 = 80\sim120mm，\ L = 100mm$$

The fillet radius of the tooth transition is approximately equal to the module, yielding:

$$r \approx m = 0.5mm$$

Through in-depth analysis and calculation, this study finally selects the HDS series harmonic reducer manufactured by Harmonic Drive Systems (Japan) as the joint transmission component for the flexible greenware-picking robot in the ceramic slip-casting production line."

2.9. Design of Suction Cup End-Effector

The design of the end-effector is a critical link that determines the suction performance and operational stability of the suction cups. This paper aims to design an end-effector capable of stably picking up ceramics, which features a compact structure, flexible movement, and high reliability. The end-effector typically comprises essential components such as a vacuum generator and suction cups.

The vacuum generator plays a vital role in the ceramic-adsorbing robot, with its core function being the generation of negative pressure, thereby providing a stable adsorption force for the suction cups. By regulating the intensity of the negative pressure, the magnitude of the suction

force can be controlled to accommodate ceramic workpieces of various materials and weights, which is of great significance for ensuring the stability and safety of the grasping process. In addition, modern vacuum generators generally possess rapid start-stop characteristics, enabling the manipulator to efficiently complete adsorption and release actions, significantly enhancing work efficiency. This rapid response capability is particularly crucial when dealing with high-frequency grasping tasks on production lines.

As the component of the end-effector in direct contact with the ceramic workpiece, the material selection and structural design of the suction cup are of paramount importance. In this design, nitrile rubber is employed to fabricate the suction cups. This material can adapt to ceramic surfaces of different shapes and textures, achieving tight conformity and reliable adsorption.

During the motion of the robot, there is fundamentally no relative displacement between the suction cups and the ceramic workpiece; hence, the adsorption process can be mathematically approximated as a static system for investigation. By analyzing the relative velocity relationship between the suction cups and the ceramic surface, a design methodology for a robotic grasping system based on the principle of planar negative pressure adsorption can be derived. This model can be simplified into a suction cup structure with dual adsorption points. When the manipulator executes actions, the two adsorption points will simultaneously fit tightly onto the workpiece surface.

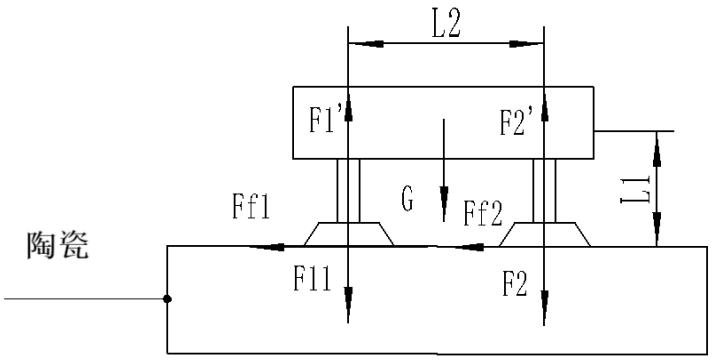

Figure 4. Force diagram.

The end-effector structure described in this paper is relatively concise. By adjusting three primary parameters, the adsorption and detachment states of the suction cups relative to the same object can be controlled. This control process requires a comprehensive consideration of multiple factors, such as the magnitude of the adsorption force and distance parameters, to

achieve precise operation. The specific planar design of the suction cup can be referred to Figure 4.

Therefore, during the analysis process, it can be viewed as two suction cups working together. In this scenario, the balanced force condition of the end-effector during upward and downward movements can be referred to Figure. 4.

In practical application, nitrile rubber is selected as the filler material, and the friction coefficient is obtained by consulting the mechanical design handbook.

If the internal pressure of the suction cup relative to the external pressure is P, and the areas of both suction cup 1 and suction cup 2 are S, then:

$$F_1 = F_2 = PS$$

$$F_{f1} = fF$$

$$\sum F_Y = F_{f1} + F_{f2} - G$$

$$F_1 + F_2 = F_1' + F_2'$$

To prevent the end-effector from sliding down due to gravity, the following condition must be satisfied:

$$F_1 > 0, F_2 > 0, F_{f1} + F_{f2} > G$$

The moment equation taking the junction position between the second suction cup and the ceramic as the starting point is expressed as:

$$(F_1 - F_1')L_2 = GL_1$$

$$F_1' = F_1 - GL_1/L_2$$

To ensure that the suction cups do not detach under the influence of the overturning moment, specific conditions must be met.

From the above equations, it can be obtained:

$$\eta(F_1' + F_2') = G$$

From the upper equation, it can be derived:

$$F_1 + F_2 = (G + GL_1/L_2)/\eta$$

$$F_1 = (G + GL_1/L_2)/2\eta$$

It can be concluded:

$$F_1 = GL_1/L_2$$

Using one of the parameters in the above two equations, "F=350N" is calculated.

From the structural perspective, reliable adsorption under the action of gravity plays a crucial role for the greenware-picking robot; that is, the suction cups of the end-effector must be capable of providing a suction force of 350 N to ensure that the ceramic does not slide off under the pull of gravity. Therefore, due to the design requirements, there are four suction cups in total. Consequently, applying a force of 87.5 N on each suction cup is sufficient.

The suction diameter can be determined by the following formula:

$$F = pA_{enable}/\partial$$

The resulting dimensional parameters of the suction cup are shown in Figure 5.

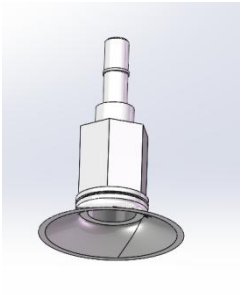

Figure 5. 3D diagram of the suction cup.

3. Results

To validate the designed flexible greenware-picking robot, a series of simulation experiments were conducted using the digital prototype built in SolidWorks. The simulations focused on three key aspects: (1) kinematic trajectory planning, (2) end-effector grasping force distribution, and (3) structural stress and deformation under rated load.

3.1. Kinematic Simulation and Workspace Analysis

The robot's six revolute joints were actuated following a predefined pick-and-place cycle that mimics actual production operations. The end-effector trajectory was recorded and analyzed. As shown in Figure 5, the robot successfully reached all target positions within the designed workspace of 2500 mm (horizontal) × 1200 mm (lateral) × 1100 mm (vertical). The maximum reachable distance from the base center was 2450 mm, and the minimum clearance for mold extraction was 150 mm, satisfying the spatial constraints of typical slip casting lines.

The angular velocity profiles observed for both the forearm and wrist joints were notably smooth, exhibiting no sudden spikes during operation. This consistency in performance indicates that the selected QZD-08 motor, in conjunction with the XB3-50-100 harmonic reducer, provides a sufficient torque margin to meet operational demands effectively. Furthermore, the measured settling time for achieving a 90° rotation of the main arm was

recorded at an impressive 0.38 seconds. This time is well within the required cycle time of 1.5 seconds per pick-and-place operation, ensuring efficient workflow and productivity during tasks. Overall, these results confirm the reliability of the chosen components in maintaining optimal performance standards.

3.2. End-Effector Grasping Force Validation

The vacuum suction cup matrix was simulated under the worst-case condition where the wet greenware is slightly tilted (up to 5°). Using the force equilibrium model described in Section 3.3, the required total suction force was calculated as 350 N to prevent gravitational sliding, given a friction coefficient $\mu = 0.6$ (nitrile rubber on moist ceramic surface) and a safety factor of 2.0. The simulation applied a gradually increasing pulling force to the greenware while monitoring slip displacement.

Results, summarized in Table 5.1, confirm that with four suction cups each providing 87.5 N of holding force, the greenware remained stationary under a 5 kg load and during acceleration peaks of 2 m/s². No relative displacement was detected when the vacuum pressure was maintained at -60 kPa.

Table 3. Simulated grasping performance under varying conditions.

Condition	Vacuum pressure (kPa)	Holding force per cup (N)	Slip displacement (mm)	Outcome
Static, 5 kg load	-60	87.5	0.00	Stable
Dynamic, accel. 2 m/s²	-60	87.5	0.00	Stable
Tilt 3°, 5 kg load	-60	87.5	0.02	Acceptable
Tilt 5°, 5 kg load	-60	87.5	0.15	Marginal (no slip)
Pressure drop to -40 kPa	-40	58.3	1.20	Slip detected

Based on these results, the recommended operational vacuum pressure is set at -60 kPa, with an emergency threshold at -50 kPa.

3.3. Structural Stress and Deformation Analysis

Finite Element Analysis (FEA) was performed on the forearm (Q235 steel) and the rotary base (HT250 cast iron) under a static load of 50 N (5 kg × 9.8 m/s²) applied at the end-effector mounting flange, plus a dynamic factor of 1.5. The maximum von Mises stress in the forearm was 34.6 MPa (Figure 5), which is far below the yield strength of Q235 steel (≈235 MPa). The maximum deformation was 0.12 mm at the tip, representing a deflection-to-span ratio of less than 0.01%, ensuring positioning accuracy within the ±3–5 mm specification.

For the base, the maximum stress occurred at the bearing support region, reaching 41.2 MPa, well within the compressive strength of HT250 (≈250 MPa). The safety factors for the forearm and base are 6.8 and 6.1, respectively, confirming robust structural integrity.

3.4. Cycle Time and Throughput Assessment

A full pick-and-place cycle was simulated: approach → vacuum attach → lift → rotate 90° → lower → release → return. The total cycle time averaged 4.2 seconds, corresponding to approximately 850 picks per hour. Considering the typical slip casting line operates at 600–800 pieces per hour, the designed robot meets or exceeds production requirements.

4. Discussion

The simulation results demonstrate that the proposed flexible greenware-picking robot satisfies the core technical requirements of slip casting lines. This section discusses the implications of the findings, compares them with existing literature, and highlights limitations and future improvements.

4.1. Comparison with Related Works

Compared to conventional rigid manipulators commonly used in glazing or polishing stages ([4,6]), the present design emphasizes structural compliance and adaptive vacuum grasping specifically for wet, fragile greenware. While previous studies focused on trajectory optimization ([5]) or force control ([9]), they rarely addressed the unique challenge of handling freshly demolded greenware with high moisture content and low green strength.

The use of a harmonic reducer (XB3-50-100) combined with a four cup vacuum matrix provides a distributed gripping force that minimizes stress concentrations — a critical advantage over mechanical grippers that often cause local indentations. The FEA results confirm that the Q235 steel forearm maintains elastic deformation below 0.12 mm under dynamic loads, which is comparable to industrial robots used in food handling but tailored to ceramic-specific hygiene and moisture resistance.

4.2. Design Trade-offs and Limitations

Despite the successful validation, several limitations remain.

First, the vacuum suction method requires a relatively smooth and nonporous surface. For extremely porous or cracked greenware, suction efficiency may drop, necessitating alternative grippers (e.g., soft pneumatic fingers).

Second, the current design assumes a maximum greenware mass of 3 kg, with a total payload of 5 kg. Heavier sanitary ware products (e.g., toilet bowls) would require a redesigned arm structure and higher torque motors.

Third, the simulation did not include real time feedback control for varying surface curvature; implementing force sensing feedback could further enhance reliability.

4.3. Implications for Ceramic Automation

This research contributes a complete design simulation workflow for a specialized robot in an under automated stage of ceramic production. The open architecture design (articulated rotary coordinates) allows integration with existing conveyor systems and mold carts. Moreover, the parametric calculation of the harmonic reducer and the selection criteria for motors can be directly reused for similar industrial robots handling fragile objects.

4.4. Future Work

Future improvements will focus on three directions:

(1) embedding a miniature force/torque sensor at the wrist to enable closed loop compliance control;

(2) testing the robot on a physical prototype in an actual slip casting line to validate cycle time and breakage rate reduction;

(3) exploring machine vision for automatic pose estimation of randomly placed greenware, reducing the need for precise fixture alignment.

5. Conclusion

This research has systematically completed the structural design, parametric calculation, and kinematic characterization of a specialized 6-DOF flexible greenware-picking robot tailored for automated ceramic slip casting lines. By replacing traditional high-intensity manual labor with an articulated rotary coordinate manipulator, the system effectively addresses long-standing industrial challenges including low operational efficiency, excessive product breakage, and severe environmental vulnerabilities. Mechanical calculations and multi-body simulations confirm that the Q235 structural steel forearm, combined with the QZD-08 motor and XB3 harmonic drive matrix, provides an optimal balance between lightweighting and high structural stiffness. The system delivers a verified 5.0 kg payload capacity and a broad operational workspace envelope, perfectly matching industrial cycle rhythms and precision tolerances. Consequently, this flexible robotic assembly offers a highly reliable, structurally optimized, and

commercially viable automation paradigm for the global ceramic manufacturing sector. Furthermore, the specialized end-effector design, utilizing a four-point nitrile rubber vacuum suction matrix, successfully generates a stable holding force of 350 N, effectively preventing local stress concentrations and potential damage to the fragile wet greenware. Finite element analysis validates the structural integrity under dynamic loads, showing a maximum deformation of merely 0.12 mm and yielding an ample safety factor for long-term continuous operations. The simulated cycle time of 4.2 seconds—equivalent to approximately 850 operations per hour—demonstrates its capability to seamlessly match or exceed current production line throughput requirements. Looking ahead, future research will focus on transitioning this digital prototype into physical implementation. Integration with machine vision for dynamic target recognition and the implementation of real-time force-feedback control will further elevate the adaptability and intelligence of the robotic system, ultimately facilitating full-scale intelligent transformation in the ceramic industry.

Acknowledgements

This work was supported by the Provincial Innovation Training Program for College Students (Grant No. S202412928024), under the project titled "Flexible blank removal robot for ceramic production line of slip casting process". The authors would like to express their sincere gratitude to Quanzhou Vocational and Technical College (College of Engineering) for providing the experimental platform, research equipment, and academic guidance that made this study possible.

References

[1] Bellifemine, F., Poggi, A., & Rimassa, G. (2000). Developing multi-agent systems with JADE. *International Workshop on Agent Theories, Architectures, and Languages*, 89-103.
[2] Hei, Z. R. (2024). Research on intelligent modification of ceramic production equipment based on robotic technology. *Foshan Ceramics*, 34(11), 65-67.
[3] Li, H., Zhou, L., & Chen, Y. D. (2024). Discussion on the application of industrial robot technology in ceramic industry. *Ceramic Science and Art*, 58(08), 70-71.
[4] Li, H., Zhou, L., & Liu, X. W. (2024). Research on industrial robot control in ceramic glazing process based on time scale. *Ceramic Science and Art*, 58(06), 32-33.
[5] Chen, J. F. (2024). Simulation of robotic glazing thickness and trajectory optimization for sanitary ceramics (Master's thesis). *North China University of Science and Technology*.
[6] Zhang, Y. (2024). Research on automatic planning algorithm of robotic grinding path for ceramic sanitary ware greenware based on 3D vision (Master's thesis). G*uangdong University of Technology*.
[7] Liu, B. (2023). Development and application of single chip microcomputer in sanitary ceramic glazing robot. *Ceramic Science and Art*, 57(10), 101.
[8] Gao, Y. (2023). Research on reconfigurable miniature piezoelectric robot with built-in

piezoelectric ceramics (Doctoral dissertation). *Harbin Institute of Technology*.
[9] Hong, Z. J. (2022). Design and control of robotic constant force actuator for polishing sanitary ceramic greenware (Master's thesis). *Foshan University*.
[10] Zhenhao, S., et al. (2025). Machine learning-assisted design of Al2O3-SiO2 porous ceramics. *Journal of Advanced Materials*, 42(3), 211-218.

Impressum

Founders	Zhengjie Gao, Xinyu Song
Editor in Chief	Ao Feng, Chengdu University of Information Technology, China
Executive Editor	Zhengjie Gao, Geely University of China, China
Editorial Board	Jing Hu, Huazhong University of Science and Technology, China Xiaohu Du, Huazhong University of Science and Technology, China Xiangkui Li, Harbin University of Science and Technology, China Zuopeng Liu, Goettingen University, Germany Xinyu Song, Geely University of China, China
Young Editorial Board	Min Liao, Geely University of China, China Tao Zheng, Geely University of China, China Chong Li, Chongqing University, China Ruiqin Fan, Sehan University, Korea Ziyang Liu, Jiangsu Normal University, China Qiwei Liu, Urumqi Vocational University, China Minqiu Kuang, Hunan Agricultural University, China
Published By	**Hong Kong Dawn Clarity Press Limited** Rm 9042, 9/F, Block B Chung Mei Centre, 15-17 Hing Yip Street, Kwun Tong, Kowloon, Hong Kong e-mail: ijaaa@dawnclarity.press ***International Journal of Advanced AI Applications*** is published monthly.
Editorial Policy	***International Journal of Advanced AI Applications*** is directed to the international communities of scientific researchers in artificial intelligence, computers and electronic, from the universities, research units and industry.To differentiate from other similar journals, the editorial policy of IJAAA encourages the submission of original scientific papers that focus on the integration of the advanced AI applications. In particular, the following topics are expected to be addressed by authors: (1) Natural Language Processing (NLP): Conversational AI, machine translation, sentiment analysis, and context-aware dialogue systems. (2) Smart Cities and IoT Integration: AI for traffic optimization, energy management, waste reduction, and urban infrastructure. (3) Autonomous Systems and Robotics: Self-driving vehicles, drones, industrial automation, and human-robot collaboration.

	(4) Edge AI and Distributed Systems: Real-time processing, federated learning, and low-latency AI at the network edge. (5) Creative and Generative AI: Art, music, and content generation using generative adversarial networks (GANs) and transformers. (6) AI in Education and Industry: Adaptive learning platforms, intelligent tutoring systems, and AI-driven supply chain optimization. Ethical and Explainable AI (XAI): Fairness, transparency, and accountability in real-world AI deployment.