Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 2212–2222
https://doi.org/10.54941/ahfe1007033

OCR-Based Quality Assessment and Auxiliary Review System for Semantic Information Extraction From Engineering Drawings

Bo Pang and Jiansong Zhang

Purdue University, School of Construction Management Technology, West Lafayette, IN 47907, USA

ABSTRACT

Optical Character Recognition (OCR) has been widely adopted to extract textual information from legacy engineering drawings, aiming to transform image-based PDF documents to semantically enriched digital models. However, the quality of drawings varies due to variations in sources and formats, which degrades the performance of OCR and lowers the accuracy of extraction results. Therefore, manual review is needed to correct OCR outputs, requiring additional time and labor. To address this issue, the authors proposed an OCR-based quality assessment method combined with an auxiliary review system to enhance both the accuracy and efficiency of textual information extraction. A set of semantic- and task-driven criteria was designed to evaluate drawing quality. A dataset of 50 bridge plans in PDF format was annotated with "high" or "low" quality labels, and the textual content was manually transcribed for OCR performance evaluation. The proposed method applied Tesseract OCR to extract textual information and automate the quality assessment process. Token-level confidence scores were computed, and drawings with an average score below 80 were classified as low-quality. In the auxiliary review system, tables detected were reconstructed, and cells with text below this confidence threshold were highlighted, enabling reviewers to focus on potentially error-prone regions. Experiments on the annotated dataset showed that the proposed method achieved a precision of 97.14% and a recall of 87.18% in classification. By excluding low-quality drawings, the precision increased by 17.84% and the recall increased by 18.96% in information extraction. Additionally, the auxiliary review system highlighted 36.81% of the cells, indicating a potential reduction of over 60% in manual review time. Overall, the proposed method provides a lightweight approach to improve OCR-based semantic information extraction from engineering drawings in terms of accuracy and review efficiency.

Keywords: Optical character recognition, Engineering drawings, Semantic information extraction, Quality assessment, Auxiliary review system

INTRODUCTION

In recent years, the Architecture, Engineering and Construction (AEC) industry has seen rapid development, leading to the accumulation of a large

number of design drawings. These drawings serve as key data sources and play a critical role in construction, maintenance, and renovation processes (Zhang, 2021). Traditionally, most of these drawings exist in paper hardcopy format, which poses challenges to documentation, review, version control, and collaboration across departments (Xuesong *et al.*, 2025).

In response, Building Information Modeling (BIM) has been applied as a key digitalization method to support the transformation from 2D drawings to digital models. Compared with paper drawings, BIM models enable better dataflow, thus significantly reducing time and communication costs in multiple tasks such as design management, construction planning, and cost estimation (Zhang and Yang, 2024). Therefore, transforming existing 2D design drawings into BIM-based digital models has become both valuable and urgent. However, manual modelling requires deep professional knowledge and substantial time and labor (Zhang, Zou and Dimyadi, 2021), especially during repetitive review process to minimize errors. Instead, an ideal solution is to automate the transformation process (Akanbi and Zhang, 2022a), which also aligns with the broader trend of automation in AEC domain.

In practice, design drawings contain two main types of information: geometric and semantic (Ondrejcek *et al.*, 2009). For example, a bridge drawing may include views, dimensions, and excavation details as geometric information, whereas material attributes, quantities, and component functions as semantic information. Many studies have focused on reconstructing 3D models from geometric features (Domínguez, García and Feito, 2012; Akanbi and Zhang, 2022a; 2022b). However, geometric information alone is insufficient for practical digital models. For example, element categories, material types, and functional roles can't be expressed in geometry, but these semantics are still essential for downstream applications such as compliance checking (Zhang and El-Gohary, 2016), cost estimation (Alathamneh, Collins and Azhar, 2024), and component identification (Li *et al.*, 2023).

Semantic information in engineering drawings is usually encoded in textual format, such as annotations, notes, and tables (Agossou *et al.*, 2020). To locate targeted semantics effectively, an essential premise is to extract all textual content accurately and make sure that no relevant information is omitted or misrecognized. Optical Character Recognition (OCR) technology detects characters in scanned drawings or PDF documents, and transforms them into digital format (Kang, Lee and Baek, 2019). In the AEC domain, OCR serves as the foundation of semantic information extraction. Previous studies have applied OCR to the extraction of information from 2D engineering drawings on specific datasets (Lu *et al.*, 2020; Zhao, Deng and Lai, 2021; Li and Zhang, 2025). However, extracting textual information accurately from real-world drawings is still a major challenge (Ondrejcek *et al.*, 2009). One key issue is the quality of drawings. Different sources and formats of drawings often result in issues such as low resolution, illegible text, skewed layouts, or incomplete table boundaries, which degrade OCR performance. To ensure reliability of OCR results, manual review is usually required, which is time-consuming and labor-intensive. This reliance on

manual verification indicates the need for more effective methods to assess drawing quality and locate potential OCR errors.

To address these problems, in this paper, the authors proposed an OCR-based quality assessment and auxiliary review system to improve the accuracy and efficiency of semantic information extraction from engineering drawings. A set of semantic- and task-driven criteria was established to identify the quality of drawings. The OCR-based method performed the assessment process automatically, filtering out low-quality drawings for manual extraction. In addition, an auxiliary review system that highlights error-prone contents was further introduced to reduce the manual review workload and improve the overall efficiency of semantic information extraction.

SEMANTIC INFORMATION IN ENGINEERING DRAWINGS

Engineering drawings carry rich semantic information that provides practical contexts and meanings to geometric representations. In BIM model transformation, machine-readable semantics attached to drawing elements is essential to interpret design intent and enable automation in downstream tasks (Yang *et al.*, 2020). From the format in which semantics are presented, semantic information in drawings can be categorized as implicit and explicit. Explicit semantics are directly expressed as text or structured annotations. For example, a table of data or a note specifying "Column C1: 400×400mm, Concrete C30" is explicit. The semantic information directly label an element and its properties in a human-readable way. Implicit semantics refer to information encoded through domain-specific symbols or drawing conventions, which require expert knowledge to decode (Abrantes Baracho and Valadares Cendón, 2012). For example, a particular hatch pattern indicates a material or section type (concrete, insulation, steel, etc.) by convention (Elyan, Garcia and Jayne, 2018). Such implicit semantics must be learned to interpret the drawings' meaning accurately (Maher and Rutherford, 1997), therefore a machine or a non-expert cannot intuitively decode them without additional knowledge.

Among explicit semantics, annotations are semi-structured and still require spatial context to interpret (Agossou *et al.*, 2020). For example, when a leader line connects an annotation to a specific component, the spatial linkage is required to identify the target entity and its semantic reference. In contrast, tables are highly structured and clearly associate textual labels with values, which facilitate automatic parsing and semantic interpretation (Van Daele *et al.*, 2021). Given their structured nature and important content, tables provide important sources for semantic information extraction. Therefore, this paper focuses on evaluating and improving the extraction of semantic information from tables in engineering drawings.

METHODOLOGY

The overall framework of the proposed OCR-based quality assessment and auxiliary review system is shown in Figure 1. The system takes engineering

drawings in PDF format as input and produces manually validated semantic information as output for downstream applications. First, a set of semantic- and task-driven criteria was established to define the quality of engineering drawings. Then these criteria were encoded and applied through an OCR-based assessment algorithm to evaluate the quality of drawings automatically. The performance of this algorithm was evaluated based on a dataset of bridge plans with annotations of "high" or "low" quality according to the criteria. In practice, the input PDF files are preprocessed and converted into high-resolution images. During this process, image enhancement techniques are applied, including noise reduction, contrast enhancement, and table boundary refinement. The goal is to reduce negative impacts of low physical quality such as blur, skew, and scan artifacts, and improve the readability of tables and text for the OCR engine. Then preprocessed images are passed to Tesseract OCR engine, which extracts both textual tokens and corresponding confidence scores. Cell-level and document-level confidence values are then calculated. Documents with a value lower than a pre-defined confidence threshold are marked as "low" quality and filtered out for manual extraction. Meanwhile, "high" quality documents are moved to further automated processing. Next, for each processed document, OCR engine generates a JSON format output with text, coordinates, and confidence scores. An auxiliary review interface highlights low-confidence regions, guiding reviewers to focus their attention on where errors are most likely. The final reviewed output provides manually validated semantic information that is accurate and reliable enough to be directly used in downstream AEC applications.

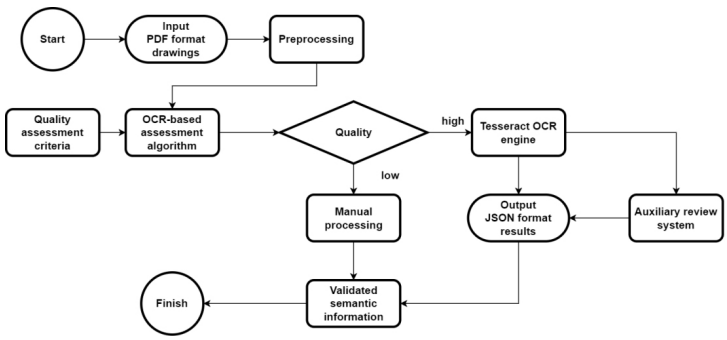


Figure 1: Framework of OCR-based quality assessment and auxiliary review system.

CRITERIA AND DATASET

To better support semantic information extraction, this paper defines a semantic- and task-driven definition of drawing quality, focusing on the presence of readable text rather than purely relying on visual or physical quality. A drawing is labelled as "high" quality only when it satisfies all the following criteria, which are derived from the practical requirements of accurate semantic information extraction:

a) No hand-written or custom fonts are present in the drawings;
b) All texts are readable without significant vagueness or blurriness;
c) The drawing contains at least one complete table with clear boundaries;
d) Background is clear, without stains, heavy shadows or scanning noise.

Drawings that fail to meet any of the above criteria are labelled as "low" quality. These criteria are designed to capture the factors that directly influence OCR performance and reliability. For example, hand-written text often leads to misrecognition, as shown in Figure 2, since Tesseract OCR and most other engines are trained on standard printed fonts and require additional specific training to deal with hand-written characters. Similarly, background noise or vague table boundaries may obscure text and degrade downstream information extraction. Figure 3 shows examples of high-quality drawings and low-quality drawings with the corresponding criterion that each fails to meet.

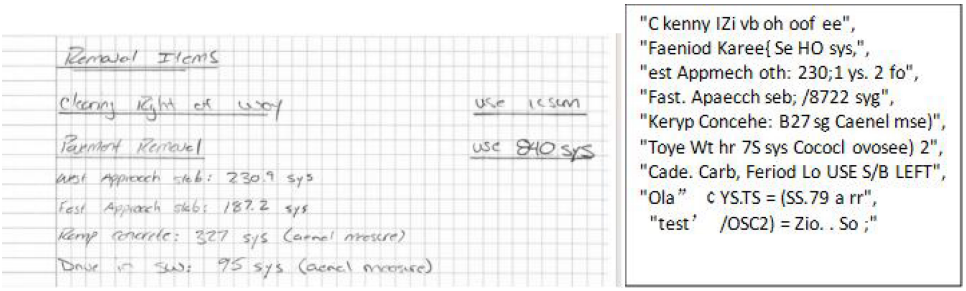
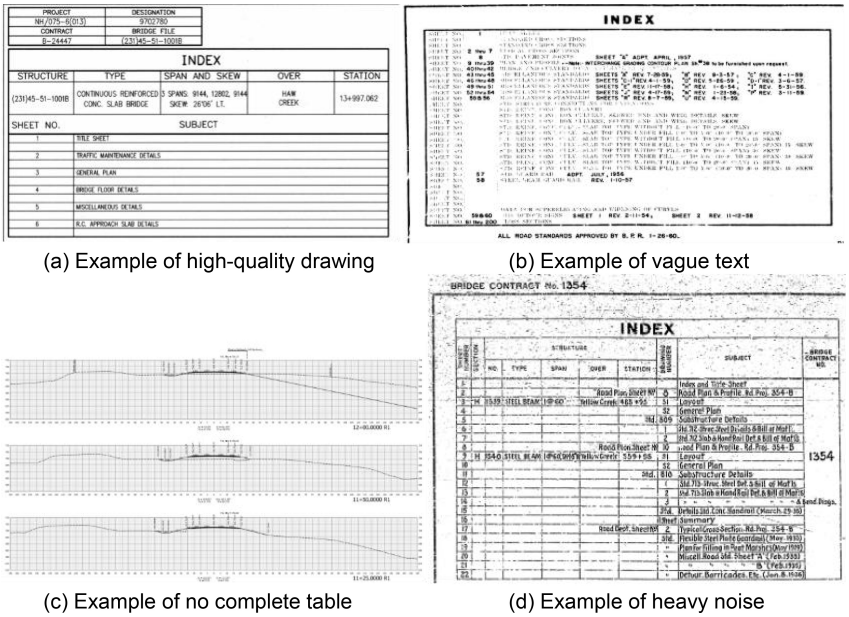


Figure 2: Example of information extraction results from hand-written text.



(a) Example of high-quality drawing
(b) Example of vague text
(c) Example of no complete table
(d) Example of heavy noise

Figure 3: Examples of typical quality issues in engineering drawings.

PREPROCESSING

The proposed system focuses on semantic information legibility rather than purely physical quality. Therefore, preprocessing is performed to mitigate the impact of physical imperfections on OCR performance. In this paper, three lightweight preprocessing techniques are innovatively combined, namely noise reduction, table boundary enhancement and text contrast enhancement. Specifically, as shown in Figure 4, (a) Gaussian Blur is applied to reduce small-scale pixel noise and smooth the background; (b) morphological operations are used to strengthen weak and broken lines, improving the structural completeness of tables; (c) Contrast Limited Adaptive Histogram Equalization (CLAHE) enhances the visibility of text edges without over-amplifying noise. These preprocessing steps collectively improve the input quality, allowing the OCR engine to concentrate on the legibility of semantic information, further supporting the semantic- and task-driven definition of drawing quality.

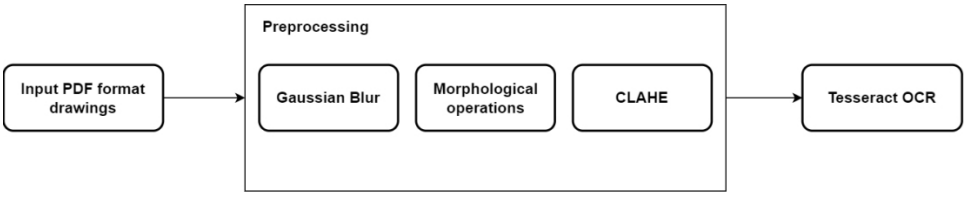


Figure 4: Steps of preprocessing.

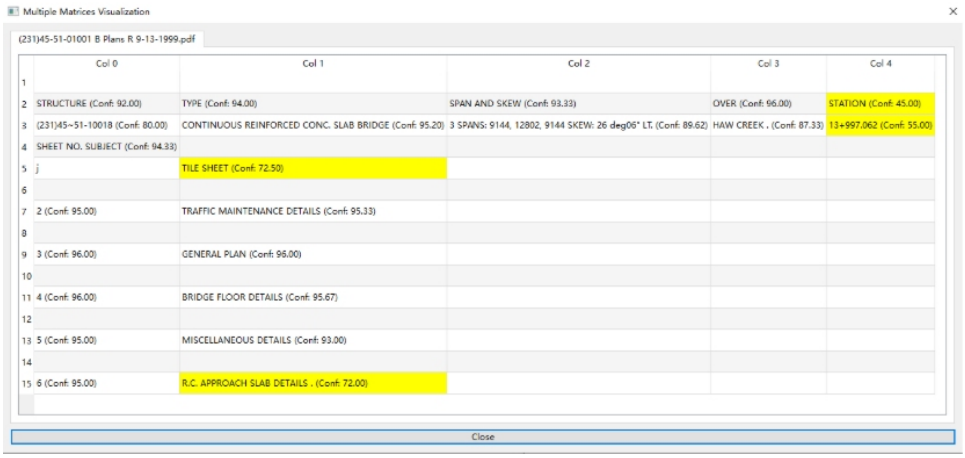
OCR-BASED ALGORITHM AND CONFIDENCE LEVEL

This system employs Tesseract OCR to extract semantic information from tables in engineering drawings. In addition to recognized text, Tesseract provides a token-level confidence score ranging from 0 to 100, indicating the reliability of recognition estimated by the model. In this paper, these scores are collected and processed for analysis of two purposes: quality assessment and confidence-based manual review.

To encode the quality identification criteria through confidence score, an OCR-based algorithm is developed to automate drawing quality assessment. The algorithm calculates the average confidence score across all recognized tokens in each drawing and table region. An empirical threshold of 80 is selected based on practical observation. Drawings that exceed this threshold are classified as "high" quality, and those below are labelled as "low" quality, indicating higher possibility of errors and requirements for detailed review. In practice, high-quality drawings can proceed to subsequent automated semantic information extraction process, whereas low-quality drawings are filtered out for manual processing.

The system also uses confidence scores to guide manual review. According to the coordinates of recognized tokens and table boundaries, the table structure is reconstructed and displayed in the user interface. Additionally, in the OCR output, each cell in the targeted table is annotated with a confidence score. The auxiliary review system highlights cells with a confidence score

below 80, indicating that detailed review is required. An example of the interface is shown in Figure 5 for manual review of semantic information extracted by OCR. The reviewers are guided to focus on areas most likely to contain recognition errors, therefore saving time and effort spent on reviewing.



	Col 0	Col 1	Col 2	Col 3	Col 4
1					
2	STRUCTURE (Conf: 92.00)	TYPE (Conf: 94.00)	SPAN AND SKEW (Conf: 93.33)	OVER (Conf: 96.00)	STATION (Conf: 45.00)
3	(231)45~51-10018 (Conf: 80.00)	CONTINUOUS REINFORCED CONC. SLAB BRIDGE (Conf: 95.20)	3 SPANS: 9144, 12802, 9144 SKEW: 26 deg06' LT. (Conf: 89.62)	HAW CREEK . (Conf: 87.33)	13+997.062 (Conf: 55.00)
4	SHEET NO. SUBJECT (Conf: 94.33)				
5	j	TILE SHEET (Conf: 72.50)			
6					
7	2 (Conf: 95.00)	TRAFFIC MAINTENANCE DETAILS (Conf: 95.33)			
8					
9	3 (Conf: 96.00)	GENERAL PLAN (Conf: 96.00)			
10					
11	4 (Conf: 96.00)	BRIDGE FLOOR DETAILS (Conf: 95.67)			
12					
13	5 (Conf: 95.00)	MISCELLANEOUS DETAILS (Conf: 93.00)			
14					
15	6 (Conf: 95.00)	R.C. APPROACH SLAB DETAILS . (Conf: 72.00)			

Figure 5: Example of auxiliary review system interface.

EXPERIMENTS AND EVALUATIONS

To evaluate the performance of proposed OCR-based quality assessment algorithm, experiments were conducted on the dataset discussed earlier. The classification results are displayed in Table 1, corresponding to a precision of 97.14% and a recall of 87.18%. The results include five false negatives (FN) and one false positive (FP). In FN cases, the drawings were mostly readable to humans, but OCR confidence scores were reduced due to minor scanning defects. This tendency shows the rigorousness of quality assessment, reducing the risk of keeping problematic drawings, thus ensuring the accuracy of extracted semantic information. The FP case involved a drawing that was of low-quality but happened to produce high confidence scores in OCR. This suggests that confidence level alone may overestimate the reliability of results in some situations. In addition, high precision indicates the system is reliable when it accepts a drawing, while moderate recall suggests a few high-quality drawings are excluded to ensure high confidence. In this context, precision is more important than recall, since minimizing errors is critical for automated semantic information extraction.

Table 1: Classification results of OCR-based quality assessment algorithm.

	Predicted High	Predicted Low
Actual high	34 (True positive)	5 (False negative)
Actual low	1 (False positive)	10 (True negative)

To evaluate the downstream benefit of the proposed quality assessment system, textual information extraction accuracy before and after applying the filter was measured. As shown in Table 2, the system achieves a precision of 72.74% and a recall of 73.63% based on 3,736 true positive words in the entire dataset. After applying the OCR-based algorithm to exclude low-quality documents, the system achieves a higher precision of 90.58% and recall of 92.59%, with 3,140 true positives from the remaining high-quality drawings. Although the total number of true positives decreased, the results confirmed that filtering out low-quality inputs improves accuracy and completeness of the semantic information extraction. This demonstrates the proposed filter is useful in real-world applications, where the quality of input data significantly affects manual review and post-processing.

Table 2: Accuracy of information extraction.

	Before	After	Improvement
TP	3736	3140	/
Precision	72.74%	90.58%	+17.84%
Recall	73.63%	92.59%	+18.96%

The proposed system also enhances the efficiency of manual review by highlighting low-confidence cells. Instead of comparing every result with ground truth, reviewers can focus on words with confidence scores below a predefined threshold. To quantify the potential time savings, the proportion of low-confidence cells was calculated over the entire dataset. As shown in Table 3, only 759 out of 2,062 cells fell below the threshold. Reviewers can perform detailed checks on just 36.81% of the content, while applying lightweight sampling or skimming on the rest, thereby reducing review workload without lowering overall accuracy.

Table 3: Proportion of highlighted cells in auxiliary review system.

Metric	Value
Cells reconstructed	2062
Cells with confidence < 80	759
Proportion	36.81%

DISCUSSION

The proposed system is designed to be lightweight, without the need for large-scale annotated training datasets. This makes it practical for integration into existing document processing pipelines. Beyond classification, the confidence-based auxiliary review system reduces manual review workload by highlighting low-confidence regions, guiding reviewers to focus on potentially erroneous segments rather than verifying every word. This improves the efficiency of human-in-the-loop workflows and enhances interpretability by providing transparent, inspectable signals regarding OCR uncertainty, which is an aspect typically missing in standard OCR outputs.

Although this paper focuses on semantic information presented in tables, the system can be extended to other forms of textual data such as annotations, schedules and part lists. By removing the dependency on table-specific rules, the system can support a wider range of semantic information extraction tasks. In addition, while the system was evaluated on bridge plans, it is generalizable and applicable to a wide range of engineering drawings, including architectural, mechanical and electrical plans.

Although the proposed methods can improve accuracy on subsequent applications and reduce time and workload spent on manual review, there are also some limitations. First, the system doesn't provide any improvement suggestions after identifying low-quality documents. As a result, manual review is still required for these cases. Second, after acquiring results, no alternative interpretations or correction suggestions are provided, leaving the correction tasks entirely to reviewers. Third, when semantic information is embedded in free-form text or visual diagrams rather than tables, the current system is insufficient. Additional modules would be required to extract and interpret such information.

Future work may develop a composite quality score that incorporates OCR confidence, image clarity, structural regularity and domain knowledge. Implementing lightweight machine learning classifiers to improve classification flexibility is another possibility. Moreover, integrating natural language processing (NLP) models could enable semi-automatic or fully automatic correction of low-confidence tokens, further reducing the need for human checking.

CONCLUSION

This paper presents a lightweight OCR-based system to assess the quality of engineering drawings and support manual review for semantic information extraction. By analysing token-level confidence scores, the system filters out low-quality drawings that are likely to degrade downstream performance, and highlights the erroneous cells in targeted tables.

Experiments on 50 bridge plans demonstrate the effectiveness of the proposed method. The filter achieved a precision of 97.14% and recall of 87.18% in classifying drawing quality, while the accuracy of semantic extraction improved by 17.84% in precision and 18.96% in recall after filtering. Additionally, the auxiliary review system can potentially reduce the manual inspection scope to approximately 36.81%, enabling more focused and efficient human validation.

The proposed system enhances OCR interpretability, reduces human workload, and can be easily integrated into practical workflows. Although current focus is on data in table format, it can be extended to other types of semantic content. Future work may explore integrating lightweight models and NLP-based correction system to further automate the review process.

ACKNOWLEDGMENT

The authors used OpenAI's ChatGPT for grammar correction and readability refinement. The authors take full responsibility for the final content and interpretations presented in this paper.

The authors would like to thank the Indiana Department of Transportation (INDOT) for providing the bridge plans used in this research. The authors also thank Bernard Nartey, Ph.D. student at Purdue University, and Inaki Garcia Barcena Garcia, undergraduate student at Purdue University, for their assistance with data annotation and experimental validation.

REFERENCES

Abrantes Baracho, R. M. and Valadares Cendón, B. (2012) "Chapter 13. An image based retrieval system for engineering drawings," in D. Rasmussen Neal (ed.) Indexing and Retrieval of Non-Text Information. DE GRUYTER SAUR, pp. 314–342.

Agossou, V. et al. (2020) "Development of a Framework to Understand Tables in Engineering Specification Documents," Applied Sciences, 10(18), p. 6182.

Akanbi, T. and Zhang, J. (2022a) "Framework for Developing IFC-Based 3D Documentation from 2D Bridge Drawings," Journal of Computing in Civil Engineering, 36(1), p. 04021031.

Akanbi, T. and Zhang, J. (2022b) "Semi-Automated Generation of 3D Bridge Models from 2D PDF Bridge Drawings," Proceedings of the Construction Research Congress 2022, pp. 1347–1354.

Alathamneh, S., Collins, W. and Azhar, S. (2024) "BIM-based quantity takeoff: Current state and future opportunities," Automation in Construction, 165, p. 105549.

Elyan, E., Garcia, C. M. and Jayne, C. (2018) "Symbols Classification in Engineering Drawings," in 2018 International Joint Conference on Neural Networks (IJCNN). 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro: IEEE, pp. 1–8.

Kang, S.-O., Lee, E.-B. and Baek, H.-K. (2019) "A Digitization and Conversion Tool for Imaged Drawings to Intelligent Piping and Instrumentation Diagrams (P&ID)," Energies, 12(13), p. 2593.

Li, H. et al. (2023) "BIM-based object mapping using invariant signatures of AEC objects," Automation in Construction, 145, p. 104616.

Li, H. and Zhang, J. (2025) "Semiautomatic IFC-Based BIM Reconstruction for As-Designed Bridges from 2D Plans Leveraging Semantic Segmentation and Enrichment," Journal of Computing in Civil Engineering, 39(6), p. 04025088.

Lu, Q. et al. (2020) "Semi-automatic geometric digital twinning for existing buildings based on images and CAD drawings," Automation in Construction, 115, p. 103183.

Maher, M. L. and Rutherford, J. H. (1997) "A model for synchronous collaborative design using CAD and database management," Research in Engineering Design, 9(2), pp. 85–98.

Ondrejcek, M., Kastner, J., Kooper, R. and Bajcsy, P. (2009) "Information Extraction from Scanned Engineering Drawings," Technical Report No. NCSA-ISDA09–001. Urbana, IL: National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign.

Xuesong, X. et al. (2025) "Associative reasoning for engineering drawings using an interactive attention mechanism," Automation in Construction, 170, p. 105942.

Yang, B. et al. (2020) "Semiautomatic Structural BIM-Model Generation Methodology Using CAD Construction Drawings," Journal of Computing in Civil Engineering, 34(3), p. 04020006.

Zhang, C., Zou, Y. and Dimyadi, J. (2021) "A Systematic Review of Automated BIM Modelling for Existing Buildings from 2D Documentation," International Symposium on Automation and Robotics in Construction (ISARC) Proceedings, 2021 Proceedings of the 38th ISARC, Dubai, UAE, pp. 220–226.
Zhang, J. and El-Gohary, N. M. (2016) "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking," Journal of Computing in Civil Engineering, 30(2), p. 04015014.
Zhang, J. and Yang, F. (2024) "Building a Bridge between Building Information Modeling and Digital Twins: Introducing Invariant Signatures of Architecture, Engineering, and Construction Objects," pp. 37–62.
Zhang, L. (2021) "Application of BIM Technology in Road Engineering Design," IOP Conference Series: Earth and Environmental Science, 760(1), p. 012009.
Zhao, Y., Deng, X. and Lai, H. (2021) "Reconstructing BIM from 2D structural drawings for existing buildings," Automation in Construction, 128, p. 103750.