Automated Image Dataset Generation Using Web Scraping and Large Multimodal Models for Construction Applications

Shrouk Gharib[1] and Osama Moselhi[2]

[1] Ph.D. Candidate, Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, QC, Canada.
[2] Professor and Director, Centre for Innovation in Construction and Infrastructure Engineering and Management (CICIEM), Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, QC, Canada.
shrouk.gharib@mail.concordia.ca, moselhi@encs.concordia.ca

Abstract –

The construction industry is undergoing a rapid transformation through the integration of digital technologies and construction. These advancements drive methods like object detection, enabling applications such as progress monitoring and material reuse to support sustainability goals. By fostering innovation, reducing material waste, and promoting circular economy principles, these technologies hold immense potential to revolutionize construction practices. However, the scarcity of dataset availability significantly hinders the development of effective machine learning (ML) applications in construction. Manual data collection remains resource-intensive, error-prone, and time-consuming, creating further barriers. This study introduces an automated methodology that combines web scraping and Large Multimodal Models (LMMs) to address these limitations. Using the Gemini Vision model, over 55,000 images were collected, including 2,691 identified as containing structural steel components—such as beams, columns, trusses, steel frames, and steel structures. Metadata generation achieved 94.8% accuracy, enabling efficient categorization through a rule-based filtering approach. The proposed methodology bridges significant gaps in the current literature by delivering a scalable, reusable framework for dataset development tailored to construction applications. By automating data acquisition and metadata generation, it enhances dataset usability for methods like object detection and applications such as progress monitoring. This research establishes a foundation for advancing AI-driven solutions in construction automation and sustainability, contributing to the industry's digital transformation.

Keywords –

Data Acquisition; Web Scraping; Construction Automation; Large Multimodal Models (LMMs); Metadata Generation; Dataset Scalability; Structural Steel Components; Automated Dataset Development.

1 Introduction

The construction industry is undergoing a transformation by the integration of digital and automated technologies. The advancements enhance efficiency, accuracy, and sustainability in tasks in tasks such as object detection for progress monitoring and material reuse planning [1–3]. These technologies hold immense potential to improve construction workflows, reduce material waste, and promote circular economy principles [4–7]. However, fully leveraging these technologies is hindered by the scarcity of high-quality domain-specific datasets, i.e. structural steel components. This limitation hampers the development, validation, and deployment of machine learning (ML) models for critical applications, including structural steel reuse, deconstruction progress monitoring, and material tracking [8–10].

Existing studies underscore the importance of datasets for progress monitoring for accurately identifying and categorizing construction materials, such as structural steel components, based on their properties, dimensions, and suitability for reuse [11–13]. However, many available datasets suffer from limited scope, inadequate diversity, and insufficient annotation quality, making them unsuitable for training robust ML models. For instance, datasets for structural steel applications often lack crucial elements such as component geometries, detailed metadata, and standardized classifications—key requirements for ML-driven solutions [14]. Moreover, manual data collection methods frequently used in this domain are labor-intensive, time-consuming, and prone to inconsistencies, further compounding the issue [15–17]. These challenges highlight the vital need for automated, scalable methodologies to develop precise and standardized datasets for ML applications.

Web scraping has emerged as a powerful tool to address these challenges by automating the collection of large datasets from diverse online sources [18] This technique has been successfully applied in domains such as healthcare and natural disaster monitoring, demonstrating its scalability and efficiency [19,20]. In construction, web scraping has shown promise in aggregating information on materials, components, and project designs, offering a viable alternative to manual data collection [21,22]. By scraping images and metadata of structural steel components from online sources, this approach can create datasets with sufficient volume and diversity to support advanced ML applications. Additionally, web scraping addresses challenges like inconsistent webpage structures and data cleaning, ensuring comprehensive and high-quality datasets [23–25].

Recent advancements in artificial intelligence, particularly in Multimodal Models derived from Large Language Models (LLMs) such as GPT-4 and Gemini Vision, further enhance the potential for dataset generation [26]. Multimodal models combine visual and textual processing capabilities, making them uniquely suited for tasks such as metadata generation and image classification [27]. These models analyze both image content and contextual text to produce descriptive metadata, enabling hierarchical organization and effective classification [28,29]. Unlike manual annotation methods, which are error-prone and inefficient, multimodal models offer unparalleled accuracy and scalability when processing large datasets [20]. Furthermore, their ability to analyze and generate structured descriptions from unstructured data makes them indispensable for research in this domain [20,30].

Despite these advancements, three critical gaps remain: (1) a lack of scalable, automated methodologies for creating domain-specific datasets in construction; (2) underutilization of multimodal models for metadata generation and hierarchical organization; and (3) insufficient exploration of the broader potential of datasets, such as their reuse in deconstruction, material tracking, or progress monitoring to support sustainability goals. Addressing these gaps is essential to advancing the field and enabling meaningful applications in construction automation, precision engineering, and sustainable practices.

This study addresses these challenges by presenting an automated methodology that integrates web scraping, multimodal models, and data organization techniques to create a scalable and reusable dataset for structural steel components. By automating the entire process—from data collection to organization—this methodology establishes a robust foundation for training ML models tailored to construction applications. Furthermore, the structured and annotated dataset produced by this study directly supports circular economy principles by facilitating material reuse, reducing waste, and optimizing resource utilization. These contributions align with sustainability goals by minimizing the environmental impact of construction practices and promoting efficient workflows, offering valuable solutions for the industry's transition toward more sustainable operations[31].

2 Objectives

This study presents an automated methodology that integrates web scraping and Large Multimodal Models (LMMs) to generate standardized, structured datasets for construction. The methodology addresses critical challenges of dataset scalability, usability, and annotation quality, with a focus on consistent and accurate component type labeling. The key objectives include: (1) develop an automated web scraping pipeline to efficiently collect construction-related images from diverse online sources; (2) utilize multimodal LMMs to generate descriptive metadata for each image, enabling precise identification and classification of construction components; (3) implement a keyword rule-based categorization system to filter and organize images into a hierarchical folder structure that represents unique construction components, improving the usability and accessibility of the dataset; and (4) preparing a dataset with unique identifiers for each image to ensure consistency, standardization, and usability for downstream applications, such as training object detection models and supporting progress monitoring and sustainability-driven practices.

3 Methodology

This study employed a systematic, multi-phase methodology to acquire, process, and categorize data related to structural steel components. Each phase was carefully designed to ensure quality, reliability, and scalability while addressing the challenges posed by diverse and large-scale datasets. The workflow depicted in Figure 1 comprises four main steps: (1) data acquisition, (2) metadata generation, (3) dataset categorization, and (4) standardization.

3.1 Phase 1: Data Acquisition

The first phase aimed to automate the acquisition of structural steel images and metadata from online repositories and websites. This required developing a robust web scraping pipeline capable of navigating diverse and often unstructured web sources.

3.1.1 Web Scraping Implementation

The web scraping pipeline was implemented in Python, incorporating the following core libraries: *BeautifulSoup* for parsing *HTML* and *XML* documents, and *Requests* for sending *HTTP* requests and retrieving webpage content, and *Urllib* for *URL* parsing and management. Figure 2 illustrates the pseudocode for the web scraping pipeline, detailing its core components; URL sanitization, dynamic crawling, image downloading, and data organization. The pipeline included:

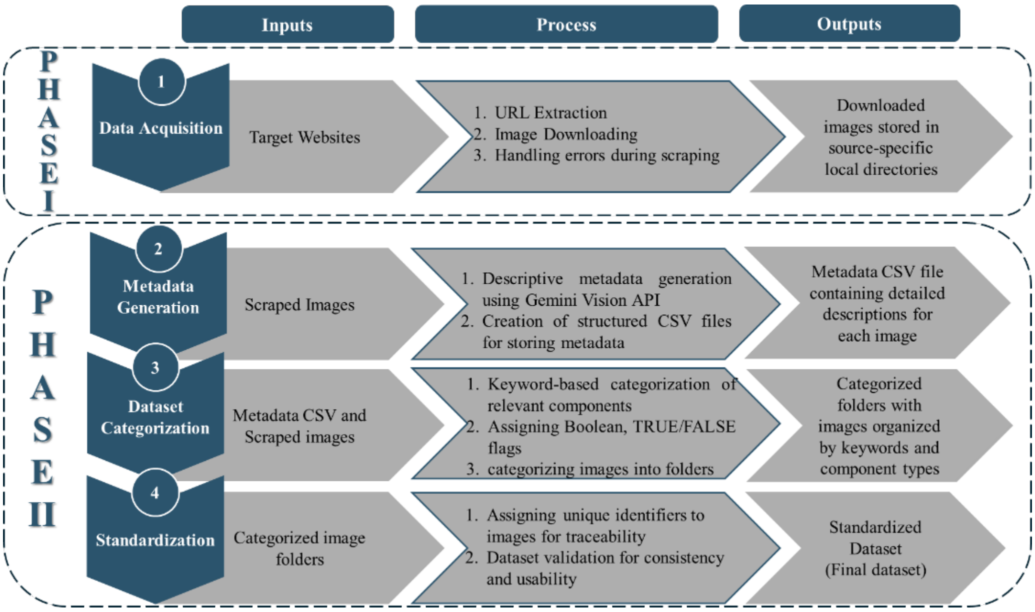

Figure 1. Workflow for Automated Dataset Generation Using Web Scraping and Large Multimodal Models

1. Sanitizing URLs and Filenames: A custom function, *sanitize_filename*, was designed to remove invalid characters from filenames and URLs, ensuring compatibility with file storage systems. This was achieved using Python's *re* module to apply regular expressions.
2. Dynamic Website Crawling: The *crawl_website* function recursively traversed web pages to retrieve images. *BeautifulSoup* was used to extract <*img*> tags, and a custom user-agent header mimicked a browser request, reducing IP blocking risks. A retry mechanism handled timeouts or rate-limit errors (HTTP 429) with randomized delays.
3. Downloading Images: image collection process relied on a targeted web scraping pipeline designed to extract domain-specific images related to structural steel. Relevant websites, including construction-focused repositories, engineering resources, and online catalogs featuring structural steel components (e.g., trusses, beams, columns, and frames), were identified and prioritized. Specific keywords such as "structural steel components," "steel trusses," "steel beams," and "steel frames" were used to guide the search and refine the scraping process. The *download_image* function saved images from extracted URLs into source-specific directories to facilitate later validation and traceability.
4. Handling Large-Scale Data**:** The pipeline could handle up to 100 pages per website and included delays between requests to avoid server flags. The *os* library was used to create a structured directory system for efficient data management.

```
IMPORT required libraries: requests, BeautifulSoup, os, re,
random
  SET base_url and target folder for image downloads
  DEFINE User-Agent for headers

DEFINE FUNCTION sanitize_filename(filename):
    REMOVE invalid characters using regex
    RETURN cleaned filename

DEFINE FUNCTION download_image(url, folder):
    SEND GET request with headers
    IF successful:
      SAVE image to the target folder with sanitized filename

DEFINE FUNCTION scrape_images(page_content, folder):
    PARSE page content using BeautifulSoup
    EXTRACT image URLs and call download_image()

DEFINE FUNCTION crawl_website(base_url, folder):
    FETCH page content
    CALL scrape_images()
    RECURSE for linked pages

DEFINE MAIN FUNCTION():
    SET base_url and target folder
    CALL crawl_website(base_url, target folder)

CALL MAIN FUNCTION() to execute the web scraping process
```

Figure 2. Pseudocode illustrating the automated web scraping pipeline.

3.1.2 Data Volume and Storage

The scraping process collected 55,926 images, organized into source-specific folders. Each image's metadata (e.g., source URL and filename) was logged for validation, ensuring efficient management and traceability in subsequent phases. Figure 1 provides a high-level overview of this process.

3.2 Phase 2: Metadata Generation, Categorization, and Standardization

The second phase aimed to enhance the collected images with detailed metadata, categorize the data, and prepare a standardized dataset for downstream applications. This phase covered Steps 2–4 of the methodology.

3.2.1 Metadata Generation

Metadata for the dataset was generated using Gemini Vision, a state-of-the-art Large Multimodal Model (LMM) designed for analyzing visual content and generating textual descriptions. The metadata generation process was implemented in Python and executed on Google Colab, leveraging its computational capabilities for efficient large-scale processing. The Gemini Vision API was used for descriptive text generation, as illustrated in Figure 3.

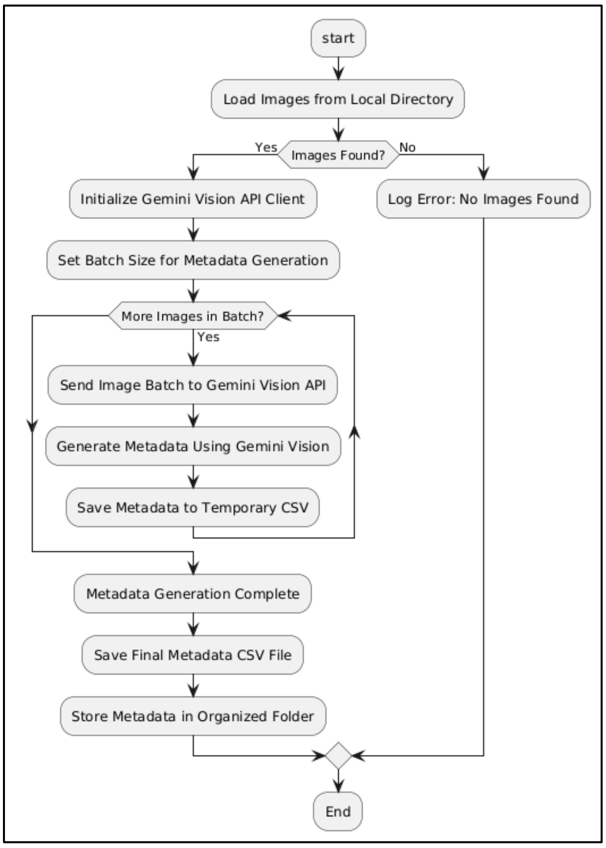

Figure 3 Workflow for metadata generation using Gemini Vision

1. Parallel Processing: To handle the large-scale dataset effectively, the *process_images_in_folder* function utilized Python's *ThreadPoolExecutor* for parallel processing. This approach distributed the workload across multiple system cores, significantly reducing processing time while ensuring scalability.
2. Error Handling: The system incorporated robust error-handling mechanisms. Issues such as corrupted image files or API timeouts were intercepted and logged with detailed error messages. Failed images were either skipped or re-queued for reprocessing, ensuring minimal data loss.
3. Metadata Validation: he metadata descriptions generated by Gemini Vision were stored in a structured Pandas DataFrame, paired with unique image IDs for traceability. To ensure accuracy, random samples (5% of the dataset) were manually reviewed, verifying their consistency with the actual visual content.

3.2.2 Metadata Example and Storage

Each metadata record included two essential fields; image ID: a unique identifier for each image, ensuring traceability, and description: a textual description generated by Gemini Vision, detailing the image's content. The final metadata was exported to a structured CSV file (*image_dataset_preparation.csv*) to facilitate seamless integration into downstream processes. This standardized file format ensured seamless integration into subsequent phases, enabling efficient accessibility and usability.

3.2.3 Rule-Based Categorization

A Python function, *extract_steel_structures*, scanned metadata descriptions for keywords indicating structural steel components (e.g., "beam," "column," "truss," "steel frame"). The categorization steps were (as illustrated in Figure 4):

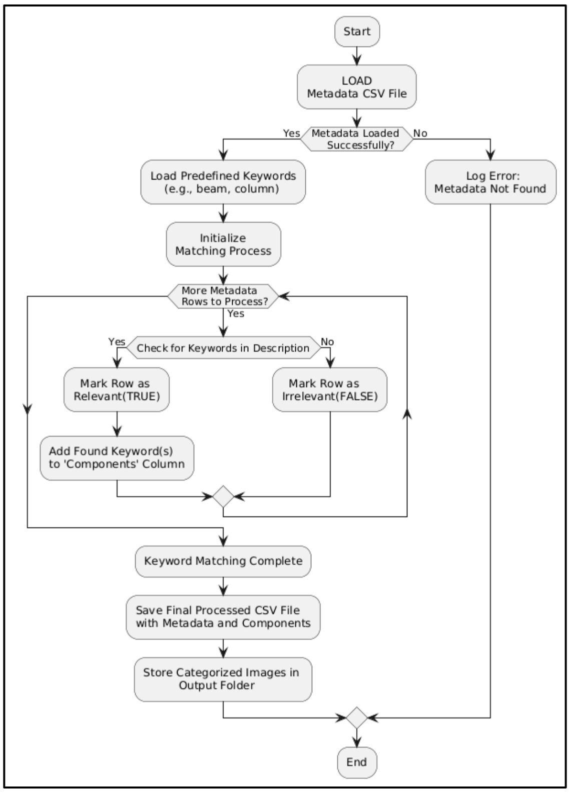

Figure 4 Keyword Matching Flowchart

1. Keyword Matching: Regular expressions were employed to identify variations in phrasing, such as "steel trusses" and "trusses of steel."
2. Field Assignment: Two fields were populated for each image: *"steel_structure"* a Boolean field *(TRUE/FALSE)* indicating the presence of structural steel components and "*components"* which is a list of matched keywords representing specific structural elements.

3.2.4 Validation and Refinement

To ensure the dataset's accuracy and reliability, a validation and refinement process was conducted. A diverse 5% sample of the dataset was manually inspected for categorization accuracy. Discrepancies, such as false positives and ambiguous descriptions, were analyzed to refine keyword matching rules. This iterative process improved classification precision and enhanced the overall quality of the dataset.

3.2.5 Final Dataset

The final dataset included 2,691 images, each labeled as TRUE for containing structural steel components. These images were categorized into the following components: beams: 1,716 images; columns: 554 images; steel frames: 169 images; trusses: 116 images; and steel structures: 506 images.

4 Results and Discussions

The proposed methodology successfully generated a comprehensive dataset of 2,691 structural steel images from an initial pool of 55,926 images retrieved from various online sources. This process demonstrated the robustness and effectiveness of the methodology in automating large-scale data acquisition, metadata generation, and dataset standardization, specifically tailored for construction applications. Each stage of the workflow addressed critical challenges, including data scarcity, inconsistent metadata, and the need for scalable solutions to support machine learning applications in the construction sector.

The web scraping process retrieved 55,926 images from diverse online sources using a pipeline optimized for dynamic website structures and large-scale data handling. Following keyword-based filtering and metadata validation, 53,235 images were excluded due to irrelevance or insufficient metadata descriptions, resulting in 2,691 images deemed relevant to structural steel components. These components were categorized into key structural elements: beams (63.8%), columns (20.6%), steel structures (18.8%), steel frames (6.3%), and trusses (4.3%). This categorization, summarized in Table 1, reflects the prevalence of these components in real-world structural steel usage, ensuring the dataset's practical applicability for construction applications.

Table 1. Categorization of Structural Steel Components in the dataset

Component type	Count	Percentage of Dataset
Beams	1,716	63.8%
Columns	554	20.6%
Steel Frames	168	6.3%.
Trusses	116	4.3%
Steel Structures	506	18.8%
Total number of images in the dataset		2,691

The metadata generation phase utilized Gemini Vision's Large Multimodal Model (LMM) to produce detailed descriptions for each image. A sample of 5% of the dataset, approximately 135 images, was manually inspected to verify metadata accuracy and categorization reliability. Observed discrepancies, including 7 false positives caused by ambiguous descriptions, informed iterative refinements of keyword matching rules and metadata prompts. These adjustments ensured alignment between generated metadata and visual content, resulting in a high accuracy rate of 94.8%. Approximately 5.2% of the metadata required manual correction due to ambiguity or incomplete details. Examples of metadata descriptions and their associated categorization are presented in Figure 5.

The metadata generation process, as summarized in Table 2, was completed in 11 hours and 42 minutes. This time refers exclusively to the metadata generation process, including image annotation and description. The workflow was implemented using Python on Google Colab, leveraging Python's *ThreadPoolExecutor* for parallel processing to efficiently handle the large-scale dataset. The computing environment included a Google Colab Pro+ subscription, providing access to an NVIDIA Tesla P100 GPU and 16 GB of RAM, which ensured sufficient computational power for metadata generation. This setup facilitated concurrent processing of multiple images, significantly reducing runtime. Dynamic crawling and retry mechanisms effectively mitigated errors caused by website structures and rate limits. This dataset provides valuable insights into structural steel components, including the unique geometries of trusses and steel frames frequently used for load distribution in roofing systems and other applications. It supports downstream machine learning applications, such as object detection models for automated progress monitoring on construction sites. Additionally, the categorized dataset holds potential for circular economy practices, enabling the identification and reuse of structural steel components during deconstruction and other sustainability-focused activities.

1	Image ID	Description	steel_structure	components
125	ecf40e32-07a5-4f44-ae42-599650b07b7e	The image shows a steel structure of a building under construction.	TRUE	steel structure
126	f935743a-4731-4c00-b64a-c19de3f474c3	The image shows a metal framework of a building under construction. There are many steel beams and girders, and the ceiling is made of corrugated metal.	TRUE	beam
127	d8e96038-154c-4474-9549-de60d88b449e	The image shows a steel structure of a building under construction. The structure consists of columns, beams, and trusses. The columns are vertical members that support the beams. The beams are horizontal members that span between the columns. The trusses are triangular members that support the roof.	TRUE	beam, column, truss
128	8e37c727-c9c4-4a5b-b422-233621258bb6	A steel frame of a container, likely used for shipping or storage.	FALSE	
129	670de234-a2c0-41dd-9a05-3374b1984fb4	The image shows a steel structure of a building under construction. There are a lot of steel beams and columns. The beams are connected to the columns with bolts and nuts. The columns are supported by the foundation. The roof is made of metal sheets.	TRUE	beam, column
130	798a0f3a-5469-40f9-b33d-11db971a0f60	The image shows a metal frame that could be used as the base for a table, bench, or other piece of furniture.	FALSE	

Figure 5 Metadata and Categorization of Structural Steel Images

Table 2 Key performance Metrics of dataset Preparation

Metric	Value	Description
Total Images Collected	55,926	Number of images retrieved from diverse online sources.
Filtered Images	2,691	Number of images retained after keyword-based filtering.
Excluded Images	53,235	Images excluded due to irrelevance or insufficient metadata.
Metadata Accuracy (For 5% of data)	94.8%	Percentage of descriptions accurately matching image content.
Processing Speed	1.33 images/ second	Average speed achieved during scraping and metadata generation.
Time Taken	11 hours, 42 minutes	Total time taken for web scraping and dataset preparation.

The methodology demonstrated significant scalability and flexibility, overcoming challenges such as dynamic webpage structures, redundant data, and linguistic inconsistencies in metadata descriptions. Iterative refinements to scraping algorithms, metadata prompts, and keyword-based categorization rules ensured a high-quality dataset, ready for diverse construction applications. These results establish a robust framework for generating high-quality datasets and pave the way for future advancements, such as integrating real-time data acquisition and expanding to other construction materials like concrete and timber. This scalable approach sets a benchmark for automated dataset generation, enhancing the adoption of AI in construction practices.

5 Summary and Concluding Remarks

This study demonstrates how web scraping and Large Multimodal Models (LMMs) can effectively create a scalable, structured dataset tailored for construction applications. By automating data collection, metadata generation, and dataset organization, the methodology overcomes inefficiencies in manual data gathering, limited dataset availability, and insufficient annotation quality. Using Gemini Vision, descriptive metadata for 2,691 structural steel images were generated with 94.8% accuracy, selected from an initial pool of over 55,000 images. These images were categorized into beams, columns, trusses, steel frames, and steel structures, enabling streamlined tasks like object detection for progress monitoring and construction automation.

Key contributions of this study include: (1) developing an automated web scraping pipeline to efficiently collect construction-related images from diverse online sources; (2) utilizing Gemini Vision for high-quality metadata generation to enable precise identification and classification of construction components; (3) implementing a keyword-based categorization system to organize and filter the dataset, using a rule-based approach with Boolean labels and hierarchical folder structures; (4) delivering a reusable dataset with standardized records and unique image identifiers to support downstream applications like progress monitoring and circular economy initiatives; and (5) establishing a scalable framework adaptable to construction materials beyond structural steel, including concrete, timber, and composites, to address broader dataset challenges.

While this study focuses on structural steel, the methodology is adaptable to other construction materials, offering potential for broader applications. However, the selection of websites might favor commonly available components, and keyword choices could introduce bias, influencing dataset composition. To minimize this bias, careful source selection and validation were implemented, along with a comprehensive keyword strategy designed to capture diverse naming conventions, ensuring balanced representation.

To ensure the robustness of the dataset. Future work should address discrepancies, such as false positives and overlooked false negatives, and explore cost-effective LMMs and open-source alternatives for wider adoption. Additional enhancements, such as real-time data acquisition from active construction sites and expanding the dataset to include diverse materials, would improve dataset usability and diversity. These developments would further AI-driven advancements in construction automation and sustainability initiatives. Overall, this study establishes a robust methodology for automated data generation and AI applications in construction. By addressing scalability, usability, and annotation challenges, it accelerates digital transformation, promotes sustainable practices, and lays the groundwork for future innovations in the built environment.

References

[1] M.A. Musarat, N. Hameed, M. Altaf, W.S. Alaloul, M. Al Salaheen, A.M. Alawag, Digital Transformation of the Construction Industry: A Review, in: 2021 International Conference on Decision Aid Sciences and Application (DASA), IEEE, 2021: pp. 897–902. https://doi.org/10.1109/DASA53625.2021.9682303.

[2] W.S. Alaloul, K.M. Alzubi, A.B. Malkawi, M. Al Salaheen, M.A. Musarat, Productivity monitoring in building construction projects: a systematic review, Engineering, Construction and Architectural Management 29 (2022) 2760–2785. https://doi.org/10.1108/ECAM-03-2021-0211.

[3] S. Gharib, O. Moselhi, A review of Computer Vision-Based Techniques for Construction Progress Monitoring, in: 2023. https://doi.org/10.22260/ISARC2023/0071.

[4] N. Zhang, J. Ye, Y. Zhong, Z. Chen, Digital Transformation in the Chinese Construction Industry: Status, Barriers, and Impact, Buildings 13 (2023) 1092. https://doi.org/10.3390/buildings13041092.

[5] C. Toșa, C.P. Paneru, A. Joudavi, A.K.M. Tarigan, Digital transformation, incentives, and pro-environmental behaviour: Assessing the uptake of sustainability in companies' transition towards circular economy, Sustain Prod Consum 47 (2024) 632–643. https://doi.org/10.1016/j.spc.2024.04.032.

[6] O. Samuelson, L. Stehn, DIGITAL TRANSFORMATION IN CONSTRUCTION – A REVIEW, Journal of Information Technology in Construction 28 (2023) 385–404. https://doi.org/10.36680/j.itcon.2023.020.

[7] H.R. Seeberg, S.M. Haakonsen, M. Luczkowski, Systematic Mapping of Circular Economy in Structural Engineering, Buildings 14 (2024) 1165. https://doi.org/10.3390/buildings14041165.

[8] P. Hradil, L. Fülöp, M. Wahlström, C. Del Castillo, The new construction products regulation: Opportunity or barrier for reused constructional steel?, in: Life-Cycle of Structures and Infrastructure Systems - Proceedings of the 8th International Symposium on Life-Cycle Civil Engineering, IALCCE 2023, 2023. https://doi.org/10.1201/9781003323020-90.

[9] D.D. Tingley, J. Allwood, Reuse of structural steel: The opportunities and challenges, European Steel Environment & Energy Congress (2014).

[10] D. Ness, J. Swift, D.C. Ranasinghe, K. Xing, V. Soebarto, Smart steel: New paradigms for the reuse of steel enabled by digital tracking and modelling, J Clean Prod 98 (2015) 292–303. https://doi.org/10.1016/j.jclepro.2014.08.055.

[11] S. Bunrit, N. Kerdprasop, K. Kerdprasop, Evaluating on the Transfer Learning of CNN Architectures to a Construction Material Image Classification Task, Int J Mach Learn Comput 9 (2019) 201–207. https://doi.org/10.18178/ijmlc.2019.9.2.787.

[12] M. Soleymani, M. Bonyani, H. Mahami, F. Nasirzadeh, Construction material classification

on imbalanced datasets using Vision Transformer architecture (ViT), (2022). https://doi.org/10.21203/rs.3.rs-1948162/v1.

[13] A. Pal, J.J. Lin, S.H. Hsieh, M. Golparvar-Fard, Automated vision-based construction progress monitoring in built environment through digital twin, Developments in the Built Environment 16 (2023). https://doi.org/10.1016/j.dibe.2023.100247.

[14] S. Kitayama, O. Iuorio, Disassembly and Reuse of Structural Members in Steel-Framed Buildings: State-of-the-Art Review of Connection Systems and Future Research Trends, Journal of Architectural Engineering 29 (2023). https://doi.org/10.1061/jaeied.aeeng-1615.

[15] Z. Liu, A Review of Research on Digital Transformation in Construction Enterprises, Academic Journal of Science and Technology 9 (2024) 300–305. https://doi.org/10.54097/00gkv045.

[16] Y. Zhang, R. Chang, W. Mao, J. Zuo, L. Liu, Y. Han, Challenges of Automating Interior Construction Progress Monitoring, J Constr Eng Manag 150 (2024). https://doi.org/10.1061/jcemd4.coeng-14637.

[17] V.K. Reja, K. Varghese, Q.P. Ha, Computer vision-based construction progress monitoring, Autom Constr 138 (2022). https://doi.org/10.1016/j.autcon.2022.104245.

[18] S. Upadhyay, V. Pant, S. Bhasin, M.K. Pattanshetti, Articulating the construction of a web scraper for massive data extraction, in: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, 2017: pp. 1–4. https://doi.org/10.1109/ICECCT.2017.8117827.

[19] S. Ganesh, A.P. Celestina, R. Jayashree, K.V. HariPriya, Web Automation in Health Care, in: 2019 IEEE International Conference on Innovations in Communication, Computing and Instrumentation (ICCI), IEEE, 2019: pp. 28–35. https://doi.org/10.1109/ICCI46240.2019.9404388.

[20] S. Rennie, M. Buchbinder, E. Juengst, L. Brinkley-Rubinstein, C. Blue, D.L. Rosen, Scraping the Web for Public Health Gains: Ethical Considerations from a ‘Big Data’ Research Project on HIV and Incarceration, Public Health Ethics 13 (2020) 111–121. https://doi.org/10.1093/phe/phaa006.

[21] S.L. Zulu, A.M. Saad, S.O. Ajayi, M. Dulaimi, M. Unuigbe, Digital leadership enactment in the construction industry: barriers undermining effective transformation, Engineering, Construction and Architectural Management 31 (2024) 4062–4078. https://doi.org/10.1108/ECAM-05-2022-0491.

[22] L. Jacques de Sousa, J. Poças Martins, L. Sanhudo, Tackling the Data Sourcing Problem in Construction Procurement Using File-Scraping Algorithms, in: The 1st International Online Conference on Buildings, MDPI, Basel Switzerland, 2023: p. 34. https://doi.org/10.3390/IOCBD2023-15190.

[23] V. Koscheyev, V. Rapgof, V. Vinogradova, Digital transformation of construction organizations, IOP Conf Ser Mater Sci Eng 497 (2019) 012010. https://doi.org/10.1088/1757-899X/497/1/012010.

[24] R.K. Sichkariuk, N. V. Kornilovska, I.A. Lurie, M.A. Voronenko, Development and implementation of effective Web Scraping methods for automated data collection and processing using Python, Informatics. Culture. Technology 1 (2024) 62–68. https://doi.org/10.15276/ict.01.2024.08.

[25] A. Omar, O. Moselhi, An Integrated Approach for Automated Acquisition of Bridge Data and Deficiency Evaluation, in: 2023. https://doi.org/10.22260/ISARC2023/0047.

[26] S. Agrawal, J. Trenkle, J. Kawale, Beyond Labels: Leveraging Deep Learning and LLMs for Content Metadata, (2023). https://doi.org/10.1145/3604915.3608883.

[27] B. Dikmans, D. Kang, A Brief Survey into the Field of Automatic Image Dataset Generation through Web Scraping and Query Expansion, Journal of Information Processing Systems 19 (2023) 602–613. https://doi.org/10.3745/JIPS.04.0288.

[28] E. Uzun, Scraping Relevant Images from Web Pages without Download, ACM Transactions on the Web 18 (2024) 1–27. https://doi.org/10.1145/3616849.

[29] E. Uzun, A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages, IEEE Access 8 (2020) 61726–61740. https://doi.org/10.1109/ACCESS.2020.2984503.

[30] A. Karpathy, L. Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions, IEEE Trans Pattern Anal Mach Intell 39 (2017) 664–676. https://doi.org/10.1109/TPAMI.2016.2598339.

[31] V. Ilin, I. Kalinov, P. Karpyshev, D. Tsetserukou, DeepScanner: a Robotic System for Automated 2D Object Dataset Collection with Annotations, (2021). http://arxiv.org/abs/2108.02555.