Layout-Aware Document Parsing with Visual-Linguistic Fusion: The DATALUX with Academic Content Service Provider

Min Chan Kim[1], Yeonkyung Kim[1], Jae Won Lee[2], Ki Hwan Kim[2], Ji Woo Kwak[2], Jae Hong Park[1]

[1] Kyung Hee University

[2] AllBIGDAT

alscks0924@khu.ac.kr, kyk9843@khu.ac.kr, jwlee@allbigdat.com, khkim@allbigdat.com, jwkwak@allbigdat.com, jaehp@khu.ac.kr

Abstract

Many organizations are increasingly relying on unstructured documents such as PDFs and scanned forms to support downstream large language model (LLM) services, including search, summarization, and recommendation. However, traditional OCR systems struggle with diverse layouts of documents, leading to frequent errors and high costs of labor. So, this study developed DATALUX - a robust document layout system that transforms unstructured documents into structured, machine-readable data suitable for automation. Built on a transformer-based detector, DATALUX incorporates several modules for layout refinement, text-visual fusion, and layer-wise optimization to improve coherence and generalization across diverse layouts. Around January 2025, we successfully deployed DATALUX into one of the largest academic content service firms (Nurimedia) in South Korea. This firm faced the challenge of extracting metadata and references from thousands of academic papers submitted in various formats. Also, the existing LLM-based tools provided unreliable results. So, they needed to process them manually, creating bottlenecks in both labor and time. However, DATALUX enabled the automatic structuring of over 100,000 research papers a year, improving extraction accuracy to over 97%, reducing costs by more than USD 185K annually, and accelerating processing speed by 8.7 times. These deployment results suggest that DATALUX enables scalable and efficient document automation in complex and high-volume environments successfully. We thus believe that our DATALUX has a significant impact on both academia and industry practices.

Introduction

Data has been treated as an important asset. Firms collect and store large volumes of business and research documents, not only for record-keeping but also as a foundation for downstream LLM services (Jeong 2023). These include summary, search, recommendation, and automated decision-making for end-users. Well-structured databases enable such services to improve efficiency for both employees and end-users (McLean, Wu, and Vercoustre 2005). Yet, much content remains in unstructured documents such as PDFs, scanned forms, etc. However, traditional OCR systems struggle with diverse layouts and complex figures, often producing recognition errors (Patel 2025). As a result, many organizations continue to rely on manual processing, such as reviewing, labeling, and indexing, even for basic information extraction. This practice is costly and time-consuming work (Hoffmann, Zettlemoyer, and Weld 2015).

For instance, one of the largest academic content service firms – Nurimedia – in South Korea said that they needed to manually label objects in PDF documents. To deliver LLM-based services such as author search, citation recommendation, and related-paper suggestion, they needed to extract metadata from more than 100,000 new papers each year. Until recently, this work was handled manually. Employees opened each PDF, copied author names, abstracts, and references one by one, and entered them into the database. The average cost of manual work was around $1.85 per paper, which amounted to about $185K annually.

So, they deployed commercial OCR and LLM-based services; but faced two problems. The recognition accuracy was particularly low on pages with figures and tables, as the system frequently broke the reading order, misdetected captions, and misclassified graphical components. The cost was also high, in some cases exceeding the expense of manual processing. So, these examples describe the problems for this research – they need a model that achieves high accuracy across diverse layouts while remaining more cost-efficient than manual processing.

To address these problems, this study developed and deployed DATALUX, a document layout analysis (DLA) system in cooperation with ALLBIGDAT, a Korean start-up firm in the document processing industry. DATALUX builds upon the transformer-based detector DINO (Zhang et al. 2022) and introduces three key enhancements. First, a Layout Fusion (Zhang et al. 2023) module integrates visual features from document images with textual features from OCR, enabling better semantic understanding of layout elements. Second, a Refine Network improves boundary detection and models relationships between layout regions through self-attention. Third, a Layer-wise Optimization (Tang et al. 2021) strategy assigns different learning rates to each layer, preserving general features while adopting higher layers to domain-specific patterns.

The system was pre-trained on the public DocLayNet (Pfitzmann et al. 2022) dataset to acquire general layout detection capability and reduce cold-start issues. Then, we fine-tuned the model on a domain-specific dataset of 62,137 annotated pages covering 20 layout classes (e.g., titles, authors' names, figures, etc.) from Nurimedia. Compared to the baseline DINO model, DATALUX achieved consistent improvements. Overall mAP increased by 2.1%, with mAP@50 and mAP@75 increasing by 1.8% and 2.0%, respectively. Detection performance for medium and large objects improved by 5.1% and 1.7%. These results demonstrate that DATALUX is both efficient and highly accurate. In practice, DATALUX outperforms commercial OCR models, providing stable recognition across diverse and irregular layouts while keeping processing costs lower than manual processing.

DATALUX has successfully deployed in many industries, including online academic content service firms, a commercial bank, and government institutions. Among them, this study focuses on Nurimedia case. After adopting DATALUX at Nurimedia, manual processing was no longer needed in the same way as before. Metadata extracted from PDFs is now automatically stored in the database. Previously, employees spent several hours per day classifying content, extracting information, and verifying results. With DATALUX, these tasks were greatly reduced. The DATALUX system achieved 97% accuracy, which also reduced the time needed for verification. As a result, work that once took a full year could now be completed in just six weeks. The cost savings are equally significant. For about 300,000 papers in the Nurimedia database, the content extraction and classification with DATALUX cost only about $23K. In the original budget, manual processing was estimated at $554K, representing a 24-fold reduction in cost. We describe the detailed measurable improvements in the Business Implication Section.

The structured database also transformed end-user services. Researchers in the Nurimedia platform can now ask questions through an AI chatbot and receive summarized answers that draw from multiple papers. Each paper is also summarized by section and equipped with a hyperlinked table of contents, which makes it easier for researchers to navigate to relevant sections. Related references are automatically suggested and formatted correctly for citation. The previously time-consuming workflow of finding papers, reading lengthy PDFs, and compiling references has now been simplified, reducing the burden of academic researchers. Unlike other reference manager tools or AI writing assistants with only partial function (i.e., reference management only), DATALUX delivers an integrated service from database search to HTML-based reading and citation management. So, this study suggests innovative use of AI with DATALUX.

Although we provide Nurimedia as the best deployment example of DATALUX, we have successfully deployed DATALUX in many other firms, such as a commercial bank, government institutions, and other production firms. For example, iM Bank, one of the commercial banks in South Korea, applied the DATALUX system to automate loan document classification. This reduced repetitive manual processing by loan officers, cutting more than 40 minutes of work per staff member each day and saving millions of dollars annually. These results demonstrate that DATALUX is not limited to a single domain. So, we believe that DATALUX successfully transforms unstructured documents into structured, machine-readable data, improving both efficiency and accuracy across diverse industries.

Related Work

Document Layout Analysis (DLA) aims to identify and classify meaningful visual and logical structures in unstructured documents. Earlier approaches relied heavily on either OCR-based rule systems or simple CNN detectors (Katti et al. 2018), which struggled to generalize diverse and irregular layouts in documents.

Recently, transformer-based visual detection models have substantially improved layout recognition accuracy and flexibility (Zhang et al. 2022). Based on the DEtection TRansformer (DETR) architecture (Carion et al. 2020), DINO (Zhang et al. 2022) can enhance training stability and convergence speed through a denoising strategy. In this approach, noisy queries are injected into the decoder alongside standard object queries, encouraging the model to make semantically grounded predictions from the early stages of training. However, DINO focuses on visual features only, which limits its ability to distinguish visually similar but semantically different elements. To address this issue, this study introduces a Layout Fusion Network that merges visual features from DINO with text and text-bounding-box information. The Layout Fusion Network allows our system

to leverage both spatial and semantic cues in layout recognition.

Then, multimodal approaches such as M2Doc (Zhang et al. 2024) and DocFormer (Appalaraju et al. 2021) have sought to overcome the limitations of vision-only models by combining both visual and textual features (e.g., BERT, RoBERTa). These enable finer discrimination among visually similar elements like body text and figure captions. However, existing multimodal architectures are typically trained in a purely end-to-end fashion, which can suffer from gradient imbalance across heterogeneous layout classes. To mitigate this problem, this study also adopted a stage- and layer-wise optimization strategy, assigning distinct learning rates to each layer to promote balanced learning throughout the network.

From an optimization standpoint, recent works have argued that uneven gradient propagation can degrade convergence and final accuracy in modalities (Zhang et al. 2022). To address this issue, we also incorporate a Refine Network, a self–attention–based module unique to our model. By doing so, we selectively emphasize important local context and correct hierarchical relationships within the fused feature space. Unlike prior DLA studies, this refinement process enables more accurate detection in documents with ambiguous or overlapping layout structures.

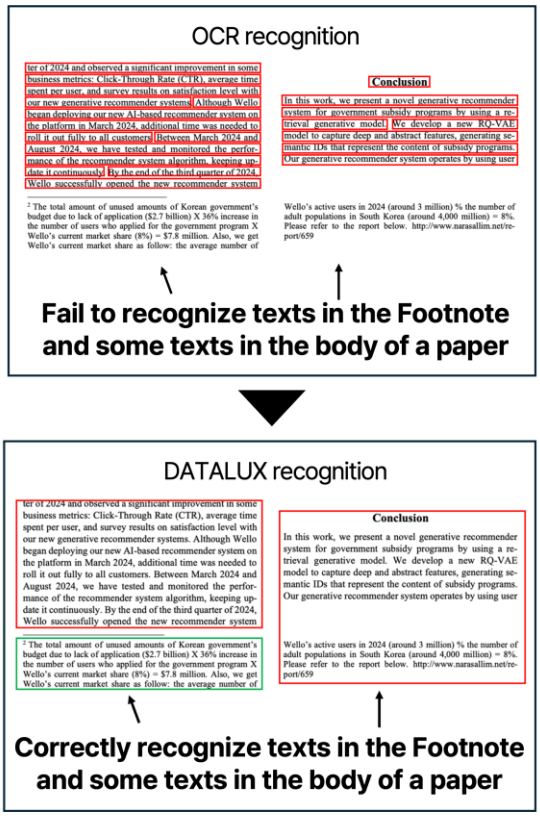

Figure 1: Recognition Results Between OCR and DATALUX.

By integrating these components (i.e., Layout Fusion, Refine Network, and layer-wise optimization) into a DINO-based backbone, our DATALUX system overcomes various limitations in prior visual and multimodal models (Xu et al. 2020). We achieve improved semantic disambiguation and detection performance across different document structures. For example, previous OCR algorithms often fail to detect texts in the footnote, while our DATALUX correctly labeled them (look at the green bounding box in Figure 1). Also, DATALUX shows a great performance in detecting texts correctly in the two-column documents. In Figure 1, although some texts in the bottom of the right column are the main body of a research paper, previous OCR couldn't recognize them. However, DATALUX robustly recognizes the texts in the bottom of the right column as the main body of the paper.

These AI approaches demonstrate how our DATALUX system extends beyond prior visual or multimodal models and provides a more reliable foundation for large-scale document analysis.

Research Dataset

We conduct our experiments with the domain-specific PDF documents provided by Nurimedia, a company established in 1997 and now recognized as the largest academic content service firm in South Korea. This firm has operated a large-scale scholarly information platform – called DBpia, serving universities, public institutions, and research organizations both domestically and internationally. Its services are subscribed to by over 1,100 institutions—including approximately 300 universities and 280 public organizations—and by more than 130 overseas research institutions. These documents consist primarily of academic papers containing diverse layout classes such as titles, authors, abstracts, tables, figures, and references.

To build a dataset for domain-specific document layout analysis, we apply a three-stage construction process. First, we analyze the structural layout patterns of the documents by identifying layout classes, spatial arrangements, and hierarchical relationships among the documents' layout elements. This process includes both the visual organization and the functional roles of each document component. Second, we examine domain-specific textual elements (i.e., terminology, formatting styles, and positional cues) and map them to their corresponding layout classes. Finally, based on these analyses, we defined 20 domain-relevant layout classes (e.g., titles, authors' names, references, figures, etc.) and performed detailed manual annotation under a strict quality assurance process to ensure label consistency and minimize errors.

So, we collected approximately 700 academic papers written in both Korean and English, covering a wide range

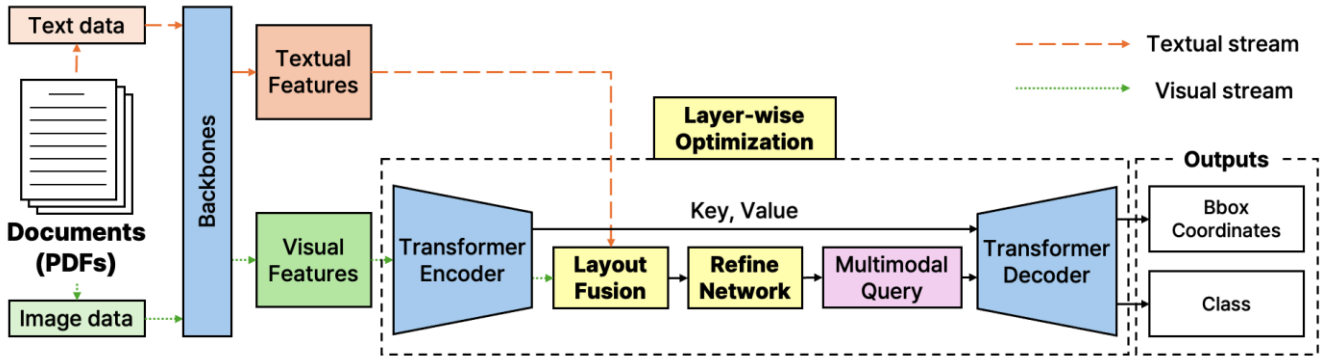

Figure 2: The overall architecture of DATALUX.

of research domains. Based on these documents, we constructed a large-scale annotated dataset containing 62,137 layout elements. The dataset was divided into training, validation, and test subsets at an 8:1:1 ratio. Each layout element was labeled according to its class, such as title, author, abstract, table, caption, or reference. This dataset served as the foundation for training and evaluating our proposed layout detection model. The detailed statistics of these layout elements can be found in Table A-1 in the Appendix Section.

Through this process, the dataset not only improves the detection performance of layout analysis models but also enables more effective learning by providing well-defined, domain-specific layout classes. By embedding domain insights into the dataset, our annotations can enhance the robustness against variations in new documents. This labeling strategy allows DATALUX to reliably transform unstructured documents into structured, machine-readable formats. It also generates accurate metadata for downstream tasks such as search, indexing, and LLM-based question answering.

Methodology

We propose **DATALUX**, an advanced layout detection model by incorporating a visual-linguistic fusion approach (e.g., M2DOC). The model combines both visual layout information and textual context to improve the identification of layout classes and the metadata quality. An overview of our model architecture is presented in Figure 2.

Base Detector: DINO

As the backbone of our model, we adopt DINO. Like DETR (Carion et al. 2020), DINO consists of three key processes: Mixed Query Selection, Look Forward Twice, and Contrastive Denoising Training (CDN). DINO offers fast and stable training even on large-scale datasets, while maintaining a relatively simple architecture (Zhang et al. 2022). Its robustness and efficiency make it suitable for business applications, and thus, we use it as the foundation of DATALUX.

Layout Fusion

While DINO relies on visual representations, we integrate a multimodal fusion mechanism to enhance its layout detection capability. DATALUX incorporates Layout Fusion, combining visual and textual features to improve detection precision.

First, the visual and textual features are extracted using ResNet (He et al. 2016) and RoBERTa (Liu et al. 2019). Visual features are fed into the DINO transformer encoder. For each predicted box r_j, overlapping OCR-derived text segments b_i are identified by $\mathrm{IoU}_{i,j} = |r_j \cap b_i| \,/\, |r_j \cup b_i|$. A text region is relevant if $\mathrm{IoU}_{i,j} > \mathrm{IoU}_{threshold}$. We then define the index set $J_j = \{i = \{1, \ldots, N\} \mid \mathrm{IoU}_{i,j} > \mathrm{IoU}_{threshold}\}$. Aggregate textual features $\mathrm{T} = (T_1, \ldots, T_N) \in \mathbb{R}^{d \times N}$ indexed by J_j yield block-level textual features. A projection function Γ maps the aggregated embedding to a fixed dimension, producing $E_j = \Gamma(\mathrm{T} \cdot J_j)$.

Finally, this block-level text embedding is injected into the corresponding decoder content query via gated additive fusion: $\mathrm{Query}_j = \mathrm{Query}_j + \lambda_1 E_j$, where λ_1 controls the contribution of textual information. By enriching queries with semantic signals from text, Layout Fusion improves the expressiveness of layout elements, leading to better region localization and more accurate classification of structural categories. The resulting multimodal queries are then used as inputs to the transformer decoder.

Refine Network

To refine the multimodal representation from Layout Fusion, we design a Refine Network by balancing the combined visual and textual representation. In Layout Fusion, the two modalities are merged through element-wise addition, which may not fully integrate their distinct properties. The Refine Network addresses this problem by blending complementary aspects of both modalities, emphasizing fine-grained details around layout boundaries. With a self-atten

Class	Base (DINO)						Ours (DATALUX)					
	mAP	mAP @50	mAP @75	mAP @s	mAP @m	mAP @l	mAP	mAP @50	mAP @75	mAP @s	mAP @m	mAP @l
Title-english	1.0	1.0	1.0	-	-	1.0	0.993	1.0	1.0	-	-	0.993
Title-korean	1.0	1.0	1.0	-	-	1.0	1.0	1.0	1.0	-	-	1.0
Author-english	0.916	1.0	1.0	-	-	0.927	0.953	1.0	1.0	-	0.9	0.968
Author-korean	1.0	1.0	1.0	-	1.0	1.0	0.958	1.0	1.0	-	0.978	0.993
Abstract-english	1.0	1.0	1.0	-	-	1.0	1.0	1.0	1.0	-	-	1.0
Abstract-korean	1.0	1.0	1.0	-	-	1.0	1.0	1.0	1.0	-	-	1.0
Caption-english	0.91	0.993	0.954	-	0.866	0.933	0.946	1.0	1.0	-	0.9	0.958
Caption-korean	0.943	0.977	0.977	-	0.934	0.949	0.934	0.992	0.982	-	0.958	0.92
Keyword	0.987	1.0	1.0	-	-	0.987	1.0	1.0	1.0	-	-	1.0
Section-header	0.962	0.988	0.988	-	0.958	0.968	0.952	0.99	0.99	-	0.953	0.957
Table	1.0	1.0	1.0	-	-	1.0	0.987	0.99	0.99	-	-	0.987
Figure	0.934	0.955	0.946	-	-	0.934	0.953	0.98	0.966	-	-	0.953
List-item	0.876	0.928	0.928	-	-	0.893	0.86	0.894	0.894	-	0.785	0.87
Equation	0.898	0.923	0.923	-	-	0.909	0.94	0.948	0.947	-	-	0.941
Text	0.895	0.914	0.902	-	0.745	0.911	0.974	0.986	0.985	0.8	0.942	0.98
Page-header	0.978	1.0	1.0	-	0.954	0.993	0.975	1.0	1.0	-	0.974	0.978
Page-footer	0.978	1.0	1.0	-	0.984	0.955	0.959	1.0	1.0	-	0.964	0.936
Page-number	0.901	1.0	1.0	0.901	-	-	0.885	1.0	0.991	0.886	-	-
Footnote	0.868	0.938	0.938	-	0.845	0.92	0.929	0.985	0.985	-	0.951	0.916
Reference	0.783	0.795	0.784	-	0.925	0.777	0.995	0.999	0.999	-	0.935	0.998
Total	0.941	0.971	0.967	0.901	0.886	0.950	**0.961**	**0.988**	**0.986**	0.843	**0.931**	**0.966**

Table 1: Performance Comparison with DINO.

tion mechanism, the model selectively focuses on structurally important regions and adjusts hierarchical cues, enhancing precision and consistency in layout detection. This refinement is particularly effective in academic documents with various boundaries and visually similar categories, such as figure captions below tables, reference lists next to footnotes, or multi-column text blocks with minimal spacing, which baseline detectors often misclassify.

Layer-wise Optimization with Fine-Tuning

To further optimize our model's performance, we apply two fine-tuning strategies. First, we employ a Layer-wise Optimization (Tang et al. 2021), which assigns learning rates individually to each layer. This strategy is particularly effective because different layers capture different levels of abstraction, and tuning them separately stabilizes training and prevents overfitting. Second, to mitigate cold-start issues, we perform transfer learning with the DocLayNet dataset (Pfitzmann et al. 2022), providing a strong initialization for domain-specific data.

With these strategies, DATALUX achieves better recognition of unstructured and visually similar layout elements than standard DINO-based detectors, enabling more accurate and consistent detection without manual intervention.

Results

We evaluate DATALUX's performance with the standard mean Average Precision (mAP) metric, which computes the average of class-wise Average Precision (AP) scores. Higher mAP values indicate better detection performance. Additionally, we report mAP at different IoU thresholds (mAP@50, mAP@75) and across object sizes (mAP@s, mAP@m, mAP@l) to assess detection qualities under various conditions.

Compared to the baseline model (DINO here), DATALUX achieves consistent improvements across most metrics. Specifically, overall mAP improved by approximately 2.1%, while mAP@50 and mAP@75 increased by 1.8% and 2.0%, respectively. Detection performance for medium and large objects also improved significantly (5.1% and 1.7%, respectively). Detailed results are presented in Table 1.

Also, we conduct additional experiments to evaluate DATALUX against other object detection models, including R-CNN–based and DETR–based approaches. Across these additional baselines, DATALUX consistently achieves higher mAP scores and demonstrates greater robustness in identifying fine-grained layout classes. For example, in our dataset, reference lists and body texts often appear visually similar—dense lines of small-sized text in paragraph

form—yet serve entirely different semantic roles. Previous systems often failed to distinguish between them, which forced users to manually separate references from body text, leading to unnecessary labor and time consumption. However, DATALUX can successfully identify them as a reference section.

We also compare the training efficiency of DATALUX with that of the DETR baseline by analyzing mAP over training epochs. The DATALUX reaches its peak accuracy in substantially fewer epochs, indicating faster convergence and stable performance.

Based on these results, DATALUX demonstrates more accurate and efficient performance than existing approaches. So, DATALUX can generate high-quality metadata more quickly and efficiently, making it particularly effective for downstream tasks such as HTML-based viewing, LLM-based services, and an internal search engine.

Deployment

Among the various deployment cases, this study focuses on Nurimedia as our best deployment practice. We successfully deployed DATALUX on the DBpia, an academic content platform, in January 2025. The Nurimedia launched DBpia in 2000. Now, the DBpia is the largest full-text academic portal in South Korea, indexing approximately 1,138 domestic scholarly journals and over 890,000 articles, with a total corpus exceeding one million papers and monthly additions of more than 10,000 new items. The engineering teams from ALLBIGDAT and Nurimedia are responsible for operating the deployed system in an on-premises environment, monitoring model performance, and updating both the model parameters and reference datasets to sustain accuracy over time. They tried to retrain the model every month. However, they learned that retraining the model monthly is not efficient, as the performance didn't change at all. So, they decided to retrain the model when a new document layout form is observed in the database. Upon detection, samples are collected, labeled, and used for fine-tuning so that the system can adapt to the new format without disrupting existing operations.

Figure 3 illustrates the details of the deployment process of our document layout system. The workflow starts when a new academic paper is uploaded to the DBpia's journal database. The raw PDF is stored securely, then DATALUX extracts its text, metadata, and page images. The layout detection engine identifies each structural element (e.g., title, author names in Korean and English, abstract, keywords, section headers, tables, figures, and references) and creates structured outputs in JSON format. These results are connected to DBpia's internal search engine and LLM-based services, enabling features such as an interactive HTML

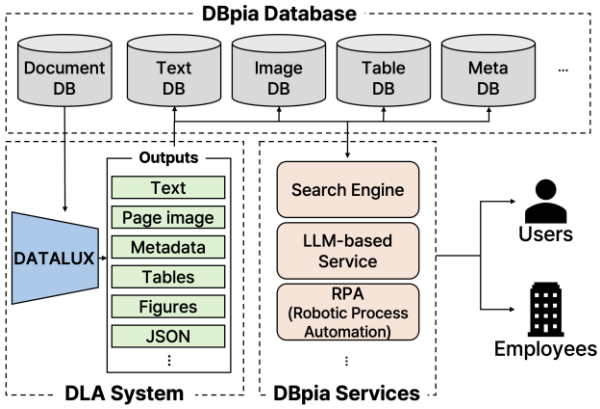

Figure 3: Deployment process in DBpia.

viewer, a clickable table of contents, and precise full-text search.

Before DATALUX, staff at Nurimedia processed every paper manually. They opened the PDF, scrolled through pages, copied text into separate fields, and retyped every reference. OCR tools were available, but often misread content, making it difficult to search later. One manager recalled that handling a 12-page paper with three tables, two figures, and 25 references could take more than 20 minutes.

This manual process also limited the services that Nurimedia could provide. The company aimed to offer LLM-based search features, such as recommending related papers or showing a particular author's other works, but these required not only accurate metadata but also reliable recognition of full-text content. For staff, categorizing the classes of similar papers was a constant challenge. They wanted to organize research by topic and recommend related studies on the website. However, the metadata provided by the authors was inconsistent. Some papers on neural networks were labeled only as "network model" or "network approach," without the word "neural." Others had keywords entered simply as "network." So, if staff search for "network", the results are mixed, very different areas such as Bayesian networks, wireless sensor networks, social network analysis, and computer networking.

However, after deploying DATALUX, Nurimedia successfully transformed such workflows. Once a PDF is uploaded, the DATALUX system detects and extracts both text and layout elements across 20 classes (e.g., figures, tables, captions, and footnotes) correctly. Metadata and content are automatically stored in the database with an accuracy of 97 percent. Figure 4 illustrates how each layout element is recognized with bounding boxes, making the document structure clear and easy to read. Previous OCR software just recognizes title, authors' information, and abstract or the main body of paper as texts, while DATALUX classifies them into each class respectively.

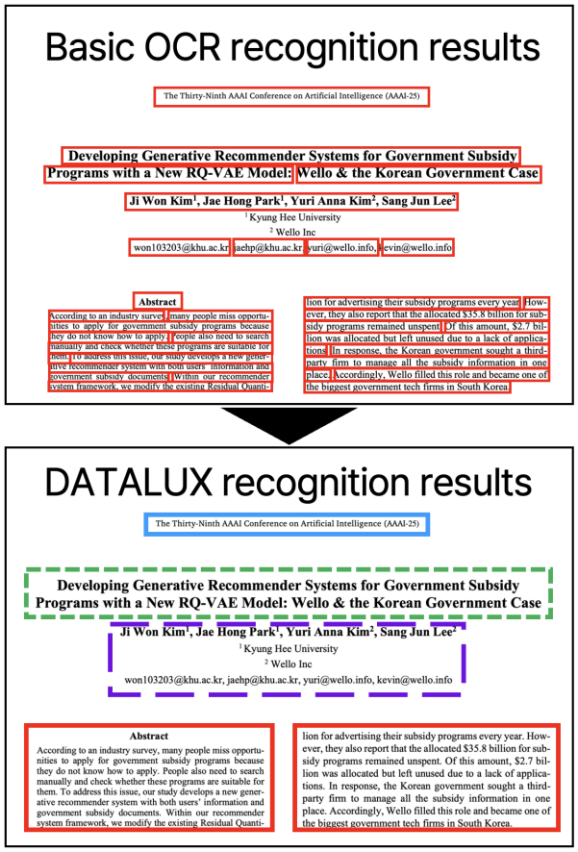

Figure 4: DATALUX recognition results with class-specific boxes, directly storable in the database.

As a result, staff no longer spend hours on repetitive data entry as search results became more accurate. Also, categorizing is automated, allowing them to quickly locate relevant papers. Staff can now focus on higher-value work such as curating the papers, producing promotional content, and developing new services.

From the end-user perspective, DATALUX at DBpia changes the way of working with academic papers significantly. Before DATALUX, the researcher relied only on a PDF viewer. Each paper had to be downloaded and opened one by one. To find specific contents, the researcher checked the table of contents, noted the page number, and then scrolled through the document to reach the section.

Even end-users needed significant time to just skim abstracts. If users needed five minutes to review one abstract, they took an hour to read the abstracts of ten papers. Later, when writing their paper, the researchers faced further difficulties. Collecting references, checking exact quotations, and formatting them in the required style all demanded time and repetitive work.

After DATALUX, the process became much smoother. The researcher can now begin with an AI-powered search interface. For example, when the researcher asks, "Tell me about document recognition algorithms", the system responds, "These algorithms detect and classify document structures such as tables, figures, and references, while recent work improves robustness across layouts." On the right-hand side of Figure 5, a panel shows related papers with titles and links, allowing the researcher to move directly from a question to further reading. Clicking one of them opens the article in an HTML viewer rather than a static PDF. In the HTML viewer, users can navigate papers more efficiently through a clickable table of contents and standard in-page search. An AI panel summarizes the paper's main points, allowing the researcher to judge its relevance quickly. Also, the "Citeasy" function suggests references related to the researcher's current manuscript and automatically formats references in the correct citation style. What used to require hours of copying, checking, and arranging bibliographic entries can now be completed in seconds.

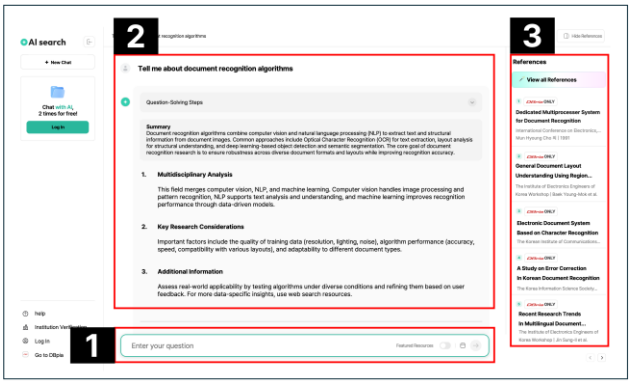

Figure 5: Example of AI-assisted paper navigation and reference retrieval in DBpia.

Compared with this integrated process, other reference management tools such as EndNote, Zotero, and Mendeley, which are widely used worldwide, mainly support storing references and formatting citations. Their AI features are limited to simple tasks such as extracting abstracts. Also, AI collaboration tools such as SciSpace provide summarization and writing support, but they do not provide direct access to academic databases or HTML-based navigation. However, DBpia, powered by DATALUX, combines these aspects in one platform. DBpia offers AI-based search and summarization of abstracts and full texts, HTML viewing with clickable navigation, and automated citation support that instantly generates references in the correct style. This integration allows researchers to move directly from discovering papers to reading and writing them. To the best of our knowledge, DBpia is the only service that supports the entire research process from finding relevant studies to writing and citing new work.

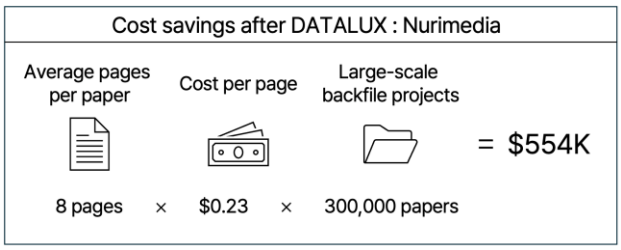

Figure 6: Cost saving after DATALUX(Nurimedia)

Business Implications

In this section, we discuss the measurable efficiency gains and economic benefits achieved through the deployment of DATALUX at Nurimedia.

At Nurimedia, DATALUX processed more than 100,000 academic PDFs in one year. According to the Korean Ministry of Employment and Labor (2024), the average hourly wage of a full-time worker was $19.93. As noted earlier, one manager recalled that "processing a 12-page paper with three tables, two figures, and 25 references could take more than 20 minutes." In practice, however, the average time per page was conservatively estimated at 40 seconds for metadata extraction. Based on this assumption, the manual per-page cost is $19.93 ÷ 3600 × 40 = $0.23. With an average length of eight pages, the per-paper cost is $0.23 × 8 = $1.85. For 100,000 papers, the annual manual budget amounts to $185,000. For a backfile of 300,000 papers, the projected cost was $554,000 (Figure 6). After DATALUX, the same project cost $23,000, equivalent to $0.076 per paper, a 24× reduction. Processing speed improved 8.7 times, calculated as 52 weeks ÷ 6 weeks ≈ 8.7, so a one-year task could be completed in six weeks. Accuracy exceeded 97 percent, and manual tagging was reduced by 95 percent.

According to the Cochrane Handbook for Systematic Reviews (Higgins et al. 2020), researchers typically spend about five minutes per abstract when screening papers. A manager at DBpia also confirms that reviewing ten abstracts often requires about 10 × 5 = 50 minutes. After DATALUX, abstracts are automatically summarized and filtered. Instead of showing all ten abstracts, the DATALUX system highlights the three or four most relevant ones. Also, DATALUX provides a summary of the most relevant abstracts of about 100–150 words, which takes about 1.5 minutes to read (Marc Brysbaert 2019). Reviewing three papers, therefore, takes 3 × 1.5 = 4.5 minutes, and four papers 6 minutes at most. This reduces the time from 50 minutes to about 5 minutes, allowing users to identify key studies much more efficiently.

The system has also been deployed in other industries. There is another deployment example – iM Bank, a commercial bank in South Korea. Previously, bank staff manually categorized loan application documents, scanning and uploading them after business hours before processing them through an OCR system. Each application took about five minutes, leading to an annual labor cost of around $6 million for an average of 8,355 daily cases. With DATALUX, the bank reduced each officer's workload by 40 minutes per day, saving an estimated $1.7 to $2.8 million annually[1].

In addition, DATALUX has been deployed in the Korea Employment Information Service (KEIS), where it was used to streamline the processing of employment-related records while maintaining accuracy levels above 95% and significantly reducing manual processing time.

Conclusion

In this study, we propose DATALUX, a document layout analysis system by extending a DINO backbone with Layout Fusion, a Refine Network, and a layer-wise optimization strategy. By integrating visual and textual features and refining them through self-attention, our model achieves more accurate and consistent recognition of diverse layout classes. Experiments show that DATALUX outperforms not only baseline DINO but also R-CNN–based and DETR–based detectors, with clear gains in mAP and robustness across different object sizes. These results demonstrate that DATALUX can effectively convert unstructured documents into structured, machine-readable formats.

As the best case of innovative use of AI, we propose Nurimedia – the operator of DBpia. DBpia is the largest academic content platform in South Korea, hosting over 4 million full-text papers and attracting about 16 million annual visitors, giving it a near-monopoly position in the Korean academic information market.

Although we propose Nurimedia with DATALUX as an innovative use of AI case, DATALUX has been successfully deployed in other industries such as government agencies and commercial banks, where DATALUX reduces manual labor, accelerates document workflows, and improves the reliability of information management. So, we believe that the impact of DATALUX will become more significant.

Acknowledgments

This work was supported by ALLBIGDAT Inc. The ALLBIGDAT and research team specially thanks to the Nurimedia for sharing the deployment experience.

[1] For the detailed improvements, please take a look at the Appendix.

Appendices

Another Example of DATALUX

As illustrated in Figure A-1, baseline DINO detectors frequently misclassify reference sections as main body text, while DATALUX successfully identifies them as reference sections.

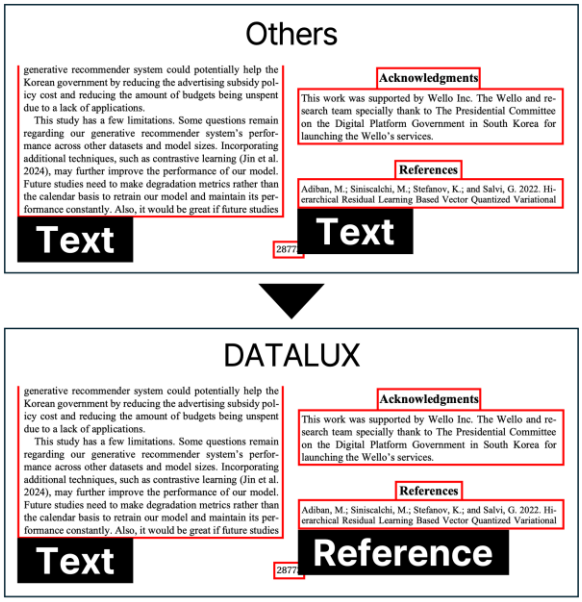

Figure A-1: DATALUX reference detection results.

Another Deployment Example – iM bank

Again, DATALUX has been deployed successfully in other industries. Another deployment case is iM Bank, a commercial bank in South Korea founded in 1976. The bank adopted DATALUX to automate the classification of loan documents across 170 branches. According to iM Bank, staff previously spent about five minutes per application checking employment, income, and property records before scanning. With an average of 8,355 applications per day, the workload is calculated as 8,355 × 5 minutes = 41,775 minutes = 696.25 hours per day. According to the Korean Ministry of Employment and Labor (2024), full-time employees worked about 236.4 days per year, which corresponds to roughly 260 business days when weekends and holidays are considered. Multiplying, the total annual workload is 696.25 × 260 = 181,025 hours. At an hourly wage of $19.93, the annual labor cost was 181,025 × $19.93 ≈ $3.6 million, and with multiple staff per branch, the total exceeded $6 million per year. After DATALUX, each officer saved about 40 minutes per day. Annual savings are calculated as 170 branches × 3–5 staff × 0.667 hours × $19.23 × 260 days = $1.7–2.8 million per year.

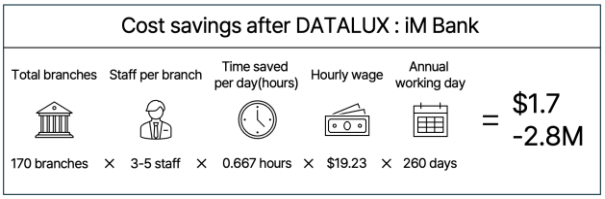

Figure A-2: Cost saving after DATALUX (iM Bank).

The Data Set Structure

Class	Train	Val	Test	**Total**
Title-english	217	31	23	271
Title-korean	348	45	43	436
Author-english	117	15	14	146
Author-korean	347	45	43	435
Abstract-english	250	36	30	316
Abstract-korean	220	31	23	274
Caption-english	1,589	207	212	2,008
Caption-korean	2,149	236	250	2,635
Keyword	216	30	27	273
Section-header	4,284	563	530	5,377
Table	778	100	112	990
Figure	2,822	336	312	3,470
List-item	1,703	285	202	2,190
Equation	648	46	84	778
Text	18,332	2,287	2,186	22,805
Page-header	2,518	331	308	3,157
Page-footer	2,017	224	238	2,479
Page-number	2,928	363	368	3,659
Footnote	702	105	78	885
Reference	7,877	759	917	9,553
Total	50,062	6,075	6,000	**62,137**

Table A-1: The detailed structure of our dataset.

References

Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. DocFormer: End-to-End Transformer for Document Understanding. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 993–1003. Los Alamitos, CA: IEEE Computer Society.

Brysbaert, M. 2019. How Many Words Do We Read per Minute? A Review and Meta-Analysis of Reading Rate. *Journal of Memory and Language* 109: 104047. Elsevier.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In Computer Vision – ECCV 2020, 213–229. Cham: Springer International Publishing.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. Los Alamitos, CA: IEEE Computer Society.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A., eds. 2020. *Cochrane Handbook for Systematic Reviews of Interventions, Version 6.1*. Cochrane Collaboration.

Hoffmann, R.; Zettlemoyer, L.; and Weld, D. S. 2015. Extreme Extraction: Only One Hour per Relation. arXiv preprint. arXiv:1506.06418

Jeong, C. 2023. A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. arXiv preprint. arXiv:2309.01105

Katti, A. R.; Reisswig, C.; Guder, C.; Brarda, S.; Bickel, S.; Höhne, J.; and Faddoul, J. B. 2018. Chargrid: Towards Understanding 2D Documents. arXiv preprint. arXiv:1809.08799

Korean Ministry of Employment and Labor. 2024. Employment Type Survey. Government of South Korea.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint. arXiv:1907.11692.

McLean, A.; Wu, M.; and Vercoustre, A.-M. 2005. Combining Structured Corporate Data and Document Content to Improve Expertise Finding. arXiv preprint. arXiv:cs/0509005

Patel, D. 2025. Comparing Traditional OCR with Generative AI-Assisted OCR: Advancements and Applications. International Journal of Science and Research 14(6): 347–351.

Pfitzmann, B.; Auer, C.; Dolfi, M.; Nassar, A. S.; and Staar, P. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 3743–3751. New York: Association for Computing Machinery.

Tang, S.; Chen, D.; Zhu, J.; Yu, S.; and Ouyang, W. 2021. Layerwise Optimization by Gradient Decomposition for Continual Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9634–9643. Los Alamitos, CA: IEEE Computer Society.

Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; and Zhou, L. 2020. LayoutLMv2: Multi-Modal Pre-Training for Visually-Rich Document Understanding. arXiv preprint. arXiv:2012.14740

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv preprint. arXiv:2203.03605.

Zhang, N.; Cheng, H.; Chen, J.; Jiang, Z.; Huang, J.; Xue, Y.; and Jin, L. 2024. M2Doc: A Multi-Modal Fusion Approach for Document Layout Analysis. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, 7233–7241. Palo Alto, CA: AAAI Press.