A Novel Dataset and Benchmark Analysis on Document Image Translation

Zhiyang Zhang[1,2], Yaping Zhang[1,2], Lu Xiang[1,2], Yang Zhao[1,2], Yu Zhou[1,3], and Chengqing Zong[1,2(✉)]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China
zhangzhiyang2020@ia.ac.cn,
{yaping.zhang,lu.xiang,yang.zhao,yzhou,cqzong}@nlpr.ia.ac.cn

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd., Beijing, China

Abstract. Document image translation (DIT) deserves more attention on account of its importance in many real-world scenarios. It is a challenging task because of the *layout degeneration* and *noisy text translation* problems caused by the optical character recognition (OCR) model. Moreover, due to the task-specific annotation, existing document image datasets usually do not support in-depth DIT analysis and model development. So, to motivate a broader investigation, this paper presents a dataset named *DITrans*, which provides fine-grained annotations for English-to-Chinese DIT task. It contains 2.8k English document images in three domains: *political report*, *scientific article* and *paper book*. Each document image has been annotated with *layout structure*, *source text* and *translation references*. Based on *DITrans*, a novel framework, which strengthens the conventional OCR-Translation cascade in *layout awareness* and *noise robustness* for better DIT, has been proposed. Furthermore, benchmark evaluations and detailed analysis based on this framework have been conducted. The evaluations and analysis results demonstrate that the dataset is very practical and can facilitate full-stack analysis and long-term research on DIT.

Keywords: Document image translation · New dataset · Benchmark · OCR · Layout structure

1 Introduction

Document image translation (DIT), aiming to perform language translation from scanned/camera document images with complex visual and layout formats, is critical for many practical applications such as translating ancient books, scientific articles and webpage screenshots, *etc.*

Y. Feng and C. Feng (Eds.): CCMT 2023, CCIS 1922, pp. 103–115, 2023.
https://doi.org/10.1007/978-981-99-7894-6_10

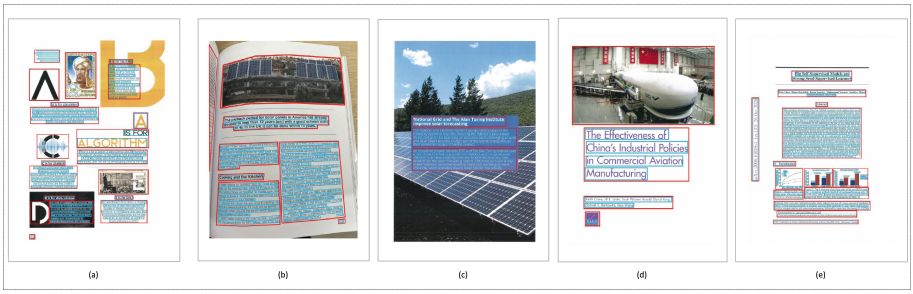

Fig. 1. *DITrans* samples. Document images are of various layouts and visual elements (glyphs, word arts, embedded figures, *etc*). Source text fragment boxes are shown in cyan. Layout blocks are shown in red. Other annotations (logical order, translation references, *etc*) are not visualized to avoid overcrowding. (Color figure online)

However, the existing method of directly joining the separately optimized optical character recognition (OCR) model and machine translation (MT) model cannot achieve optimal DIT, because of the following two problems.

- **Layout degeneration**. After being processed by OCR parser, a document image loses its layout structure and logical order, degenerating into a batch of unordered, semantically truncated text fragments. *E.g.*, the text fragments enveloped by cyan boxes in Fig. 1.
- **Noisy text translation**. The OCR output is noise-contaminated text instead of clean text as in the training phase of the MT model.

Such problems burden the translation process, making DIT much more challenging than plain text MT. Therefore, we claim the two capabilities that a superior DIT model should possess - the awareness of layout structure and the robustness to OCR noise. Nevertheless, existing document image datasets [3,6,8–11,15–18] may not support the development of such models and in-depth analysis for DIT, because their monotonic annotation is targeted for individual tasks (*i.e.*, layout analysis, logical order detection, OCR), instead of the comprehensive DIT. Therefore, to facilitate long-term research on DIT, it is inevitable to create a real-world dataset with fine-grained annotations, which support the detailed analysis of all intermediate sub-modules and innovative attempts such as layout structure utilization, OCR noise reduction, *etc*.

To this end, we have developed a novel dataset named *DITrans*, which is characterized by multiple domains and fine-grained annotations for English-to-Chinese DIT. It contains both synthetic and human-annotated high-quality annotations for document images in three domains: *political report*, *scientific article* and *paper book* (Fig. 1). Each document image has been annotated with *layout structure*, *source text* and *translation references* (Fig. 2). With these fine-grained annotations, *DITrans* is suitable for multiple document image tasks, including layout analysis, logical order detection, OCR and DIT.

Based on *DITrans*, we have developed a novel DIT framework. It is based on the conventional OCR-MT cascade and is strengthened in two aspects: 1) For **layout degeneration** problem, an additional layout analysis module is integrated with our layout-aware aggregation strategy to make the whole system aware of layout structure and logical order. 2) For **noisy text translation** problem, the pre-trained translation module is further enhanced with our modified adversarial stability fine-tuning strategy to be robust to OCR noise. Based on this framework, benchmark evaluations of various system variants are explored extensively. Furthermore, detailed analysis of the **layout structure** and **OCR noise** has been conducted. The discovery is very inspiring for new methodologies. In summary, our contributions are three-fold:

- A new dataset for English-to-Chinese DIT has been constructed. It provides fine-grained annotations including *layout structure*, *source text* and *translation references* for 2.8k document images in three domains.
- A novel DIT framework has been proposed and the performances of different system variants have been benchmarked.
- The impact of layout structure and OCR noise on DIT has been carefully studied to enlighten the proposal of new methodologies.

2 DITrans

In this section, we first give a detailed introduction to *DITrans*' fine-grained annotations and then give a cursory review of its construction workflow.

2.1 Fine-Grained Annotations

As shown in Fig. 2, the annotations can be represented as a triplet (*layout structure, source text, translation references*). Each element contains more fine-grained annotations described below.

Layout Structure. A document image is composed of multiple layout blocks (paragraph, table, figure, etc.) in a certain layout and logical order (Fig. 2 (a)). In *DITrans*, each layout block is annotated with *layout id*, *layout attribution* and *layout box*, as shown in Fig. 2 (b). With annotations of layout structure, *DITrans* is suitable for tasks including physical layout analysis and logical order detection.

- *Layout id.* It gives the logical order of a layout block. *E.g.*, A *layout id* of *4* represents the fourth reading object according to human reading order.
- *Layout attribution.* It gives a semantic label to each layout block. 14 semantic labels are defined for *DITrans*: {author-info, caption, math, image, header, footer, footnote, page-number, list-label, paragraph, reference, heading, table and unknown}.
- *Layout box.* It indicates the position of a layout block. Specifically, each layout block is enveloped by a rectangular box, with the coordinates of its upper left vertice (x_ul, y_ul) and lower right vertice (x_lr, y_lr) extracted.

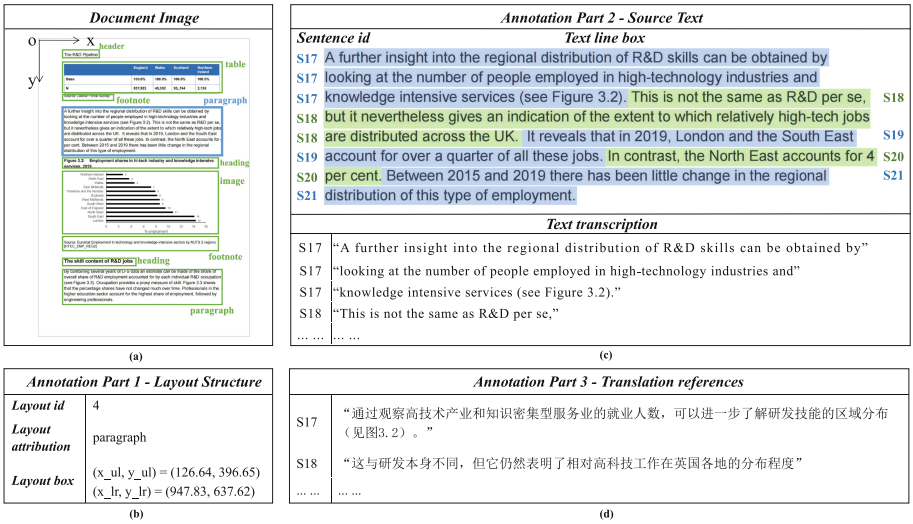

Fig. 2. An annotation example of *DITrans*. The annotation is represented as a triplet (*layout structure, source text, translation references*). Each element contains more fine-grained annotations. (a) Document image. The bounding boxes in green and blue wrap the objects to be annotated. For brevity, here we only present the fine-grained annotations for the blue box in (b) (c) and (d). (b) *Layout structure*, including *layout id*, *layout attribution* and *layout box*. (c) *Source text*, including *sentence id*, *text line box* and *transcription* of each source text fragment. (d) Chinese *translation references*, annotated at sentence level. (Color figure online)

Source Text. Source text contains *sentence id*, *text line box* and *text transcription* as shown in Fig. 2 (c).

- *Sentence id*. It indicates the reading order of a sentence. For cross-line sentences, we annotate their text fragments line by line and assign the same *sentence id* to these text fragments (Fig. 2 (c) top).
- *Text line box*. It indicates the position of a sentence. Each text fragment is enveloped by a rectangular box and the coordinates of this box are extracted to be the text line box.
- *Text transcription*. Text transcription is also annotated fragment by fragment (Fig. 2 (c) bottom), and each transcribed text fragment is aligned with its corresponding *sentence id* and *text line box*.

Translation References. As shown in Fig. 2 (d), the Chinese translation of each source sentence is given.

2.2 Dataset Construction

DITrans is composed of human-annotated data and synthetic data. The former is of high quality and the latter can be leveraged as augmented data to further improve model performance.

Human-annotated data includes three domains: *political report*, *scientific article* and *paper book*. Each domain corresponds to a unique data source: 1) *Political reports* are from the British Government Report Collection[1] and amount to 1,397 pages of 5 topics including economy, education, environment, health and technology. 2) *Scientific articles* are from arXiv and amount to 117 pages of 3 topics including machine learning, computer vision and natural language processing. 3) *Paper books* amount to 137 pages of 2 topics including opinion reviews and instructional notes. Then, the collected political reports and scientific articles were scanned and paper books were photographed for the next human annotation process. We hired 35 professional annotators for layout structure and source text and 33 translators for document translation. Translators were shown a document image with layout and source text and were required to produce correct and fluent translations in Chinese. For quality control, We hired 8 professional annotators to sample and check the annotated instances.

To further enrich *DITrans* with **synthetic data**, we automatically synthesized the complete annotations for 1,170 document images from DocBank [8]. First, we retrieved 1,170 document images belonging to the computer science domain with keyword-matching heuristics and further human inspection. Second, for each document image, words belonging to the same layout block were aggregated into a text block and segmented into sentences with an unsupervised segmentation algorithm.[2] Then, the *text line boxes* were obtained by merging adjacent word boxes. After reordering the layout blocks with the reading flow algorithm [12], the layout structure and source text annotations were produced. Finally, a commercial machine translation tool[3] was employed for sentence-level translation. This workflow extended *DITrans* with 1,170 *scientific article* document images. This part of data is referred to as *DocImg-syn*.

2.3 Dataset Statistics and Comparison

We compare *DITrans* with other widely-used document image datasets in Table 1. Statistics of *DITrans* are shown in Table 2. The two unique features of our dataset are summarized as follows.

- ***Realistic and Multi-domain***. Three domains of document images are provided, whose acquisition approaches are scanning and photographing - two common ways to obtain document images in real-world scenarios.
- ***Fine-Grained Annotations***. Fine-grained annotations are provided to make *DITrans* applicable for multiple document image tasks, such as DIT, OCR, layout analysis, logical order detection, etc.

[1] https://www.gov.uk/.
[2] http://www.nltk.org/api/nltk.tokenize.html.
[3] https://fanyi-api.baidu.com/.

Table 1. Comparison between *DITrans* and some existing document image datasets.

Dataset	Annotation	# Pages	Domain	Acquisition	Layout Block Order	Layout Box	Layout Attribution	Source Text Transcription	Translation References
PubLayNet(Zhong et al. [18])	automatic	360k	scientific article	converted from PDF		✓	✓		
DocBank(Li et al. [8])	automatic	500k	scientific article	converted from PDF		✓	✓	✓	
ReadingBank(Wang et al. [15])	automatic	500k	scientific article	converted from WORD	✓			✓	
PRImA(Antonacopoulos et al. [1])	manual	1.2k	magazine, technical article	scanning		✓	✓		
DITrans	**manual & automatic**	**2.8k**	**political report, scientific article, paper book**	**scanning & photographing**	✓	✓	✓	✓	✓

Table 2. Statistics of *DITrans*.

Domain	Annotation	# Pages	# Layout Blocks	# Source Sentences	# Words	Average #Words in Each Layout Block	Average #Words of Each Source Sentence	Average #Words of Each Translation Reference
Political Report	Human-annotated	1,397	11,980	37,691	599,000	49.82	21.43	20.91
Scientific Article	Human-annotated	117	1,357	3,978	94,990	69.25	18.93	22.03
Paper Book	Human-annotated	137	1,096	2,887	62,472	56.83	20.26	20.18
DocImg-syn	Synthetic	1,170	11,048	36,212	718,120	65.25	17.04	17.19

3 Benchmark

3.1 Layout-Aware Robust DIT Framework

The DIT task is defined as generating the translated document in logical order for a given document image. In this section, we introduce our novel layout-aware robust DIT framework - LARDIT (Fig. 3), composed of three working stages described in detail below.

Stage 1: Text Extraction and Layout Analysis. *1) Text extraction* aims at extracting the source text from a given document image I. Specifically, an OCR parser is applied on I with the resulting text fragments $F = \{f_1, f_2, ..., f_n\}$, which is typically organized in rule-based order (*e.g.*, from top to bottom). F can be further decomposed into OCR fragments text $F^{text} = \{f_1^{text}, f_2^{text}, ..., f_n^{text}\}$ and OCR fragments position $F^{pos} = \{f_1^{pos}, f_2^{pos}, ..., f_n^{pos}\}$. *2) Layout analysis* aims at parsing the layout structure from I both physically and logically, making layout blocks first detected and then arranged in logical order. **i) Physical layout analysis** returns a sequence of detected layout blocks $B^{'} = \{b_1^{'}, b_2^{'}, ..., b_m^{'}\}$ from I. For each layout block $b_i^{'}$, it determines its category label $b_i^{'ctg}$, its bounding box position $b_i^{'pos}$, and the confidence score $b_i^{'cfd}$. **ii) Logical order detection** employs the reading flow algorithm [12] to arrange the detected layout blocks $B^{'}$ into $B = \{b_1, b_2, ..., b_m\}$ with logical order instead of the original descending order of confidence score.

Stage 2: Layout-Aware Aggregation. With the extracted text fragments and layout structure, a novel mapping strategy is proposed to aggregate the text semantics $F = \{f_1, f_2, ..., f_n\}$ and visual layout $B^{pos} = \{b_1^{pos}, b_2^{pos}, ..., b_m^{pos}\}$

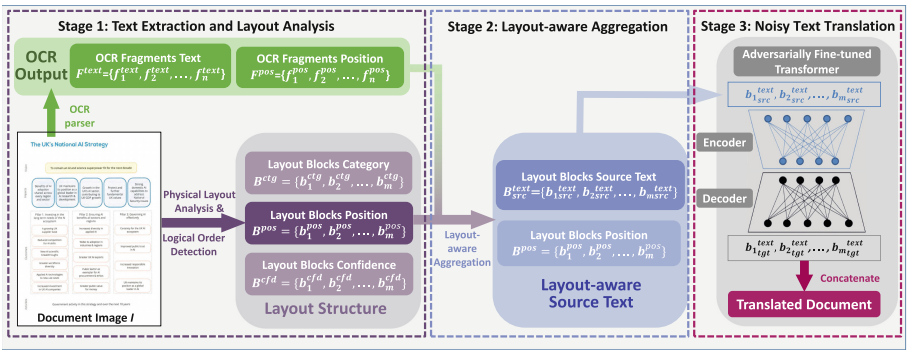

Fig. 3. Schematic diagram of our DIT framework LARDIT.

to form the well-ordered source text blocks $B_{src}^{text} = \{b_{1\ src}^{text}, b_{2\ src}^{text}, ..., b_{m\ src}^{text}\}$. Specifically, for the i_{th} element of B_{src}^{text}:

$$b_{i\ src}^{text} = concat(f_j^{text} | f_j^{pos} \in b_i^{pos}), \tag{1}$$

where $concat(\cdot)$ denotes the concatenation of text fragments from top to bottom according to their y coordinates.

Stage 3: Noisy Text Translation. To alleviate the OCR noise problem, the adversarial stability fine-tuning strategy [2] is first modified to adapt to our task and then leveraged to fine-tune the pre-trained machine translation model for better tolerance to OCR noise. The core idea is to improve the robustness of both the encoder and decoder of translation model. To this end, during the training phase, the encoder is encouraged to output similar intermediate representations for both the original and adversarial input; The decoder is guided to generate the correct output given the adversarial input or original input. In our modification, instead of manually constructing the pseudo adversarial input, which may have a distribution gap from the real-world noise, we directly treat the OCR noisy text as adversarial input. During the test phase, source text blocks in $B_{src}^{text} = \{b_{1\ src}^{text}, b_{2\ src}^{text}, ..., b_{m\ src}^{text}\}$ are sequentially translated into target text blocks $B_{tgt}^{text} = \{b_{1\ tgt}^{text}, b_{2\ tgt}^{text}, ..., b_{m\ tgt}^{text}\}$ and are finally concatenated to form a well-ordered translated document.

3.2 System Variants

We have evaluated diverse system variants based on LARDIT, each composed of different module combinations as described below.

Layout Analysis Module Variants. We build upon object detection models for physical layout analysis. Two popular object detection networks are experimented with. *a) Faster R-CNN* [13]. Its Region Proposal Network (RPN) shares the convolutional feature map with the detection network through attention

Table 3. Benchmark results of various system variants based on our framework.

System Variants		Evaluation Results											
Layout Analysis Module	TranslationModule	Political Report				Scientific Article				Paper Book			
		BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
Faster R-CNN	Transformer-PT	16.27	13.73	29.39	36.45	8.58	17.65	30.50	25.34	11.45	7.95	18.04	30.35
	Transformer-FT	23.58	15.49	33.35	44.65	16.16	19.19	33.27	34.51	13.63	9.34	23.83	33.49
	Transformer-ASF	27.41	35.78	55.25	49.17	22.01	34.48	50.22	40.28	15.87	13.32	31.65	39.12
Mask R-CNN	Transformer-PT	16.05	13.24	28.51	35.88	8.15	17.47	30.67	24.81	11.30	7.68	18.09	29.43
	Transformer-FT	23.36	16.56	33.92	43.87	16.38	18.85	33.21	34.47	13.14	9.12	22.01	33.96
	Transformer-ASF	**30.38**	**41.72**	**59.86**	**51.08**	**22.81**	**42.37**	**56.11**	**41.05**	**19.91**	**15.23**	**38.62**	**40.10**

mechanism, thereby reducing the computational overhead of region proposals. *b) Mask R-CNN* [5]. It outputs feature maps with gradually reduced resolution in multiple stages to a Feature Pyramid Network (FPN).

Translation Module Variants. Three translation models based on Transformer [14] are experimented with. *a) Transformer-PT.* Transformer-base that is pre-trained on WMT22 en-zh bilingual parallel corpus. *b) Transformer-FT.* Transformer-PT is further fine-tuned on our *DITrans* training set. *c) Transformer-ASF.* Transformer-PT is further fine-tuned on *DITrans* training set with our modified adversarial stability fine-tuning strategy.

3.3 Benchmark Results

Benchmark results are shown in Table 3. **First**, the combination of {Mask R-CNN + Transformer-ASF} is the best-performing system. Faster R-CNN and Mask R-CNN perform at roughly the same level if combined with Transformer-PT/FT. **Second**, fine-tuning the translation module improves the performance of all three domains by a large margin. This indicates a domain difference between *DITrans* and WMT22 news corpus, and fine-tuning mitigates the domain inconsistency. **Third**, our modified adversarial stability fine-tuning strategy significantly improves the translation for all three domains, demonstrating its effectiveness in addressing OCR noise.

4 Analysis

In this section, we conduct detailed analysis from perspectives of **OCR noise** and **layout structure** to enlighten new methodologies for DIT's **noisy text translation** and **layout degeneration** problems.

4.1 Analysis of OCR Noise

Overall Impact. As shown in Table 4, DIT falls behind the plain text translation of ground truth with a large margin for all three domains: 11.51, 11.11 and 9.31 BLEU declines for *political report*, *scientific article* and *paper book*, respectively. The main reason lies in the OCR noise, which causes a distribution shift to the translation module from clean text during training to noisy text during testing.

Table 4. Comparison between translating the OCR noisy source text and translating ground truth clean source text.

Noisy Text or Ground Truth	Political Report				Scientific Article				Paper Book			
	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
Faster R-CNN + OCR noisy text + Transformer-FT	23.58	15.49	33.35	44.65	16.16	19.19	33.27	34.51	13.63	9.34	23.83	33.49
Faster R-CNN + ground truth + Transformer-FT	**35.09**	**48.20**	**63.98**	**56.26**	**27.27**	**46.75**	**59.89**	**46.00**	**22.94**	**22.71**	**51.33**	**45.83**

Table 5. CER of OCR and its impact on translation. BLEU_GT and BLEU_OCR correspond to ground truth translation and OCR noisy sentence translation, respectively.

	CER↓	BLEU_GT↑	BLEU_OCR↑	Δ BLEU
Political Report	13.82%	39.5	24.9	14.6
Scientific Article	19.04%	30.7	16.4	14.3

Detailed Analysis. We further set a simplified version of the DIT task by shielding other variables (layout, sentence segmentation, *etc.*) to concentrate on the sentence-level fine-grained analysis of OCR noise. Specifically, we extract each sentence's image patch with the gold-standard line box and apply the OCR-MT procedure to get its noisy source sentence and translation.

First, the OCR character error rate (CER) and BLEU decline caused by OCR noise are shown in Table 5. *The CER of political report and scientific article are 13.82% and 19.04%, respectively, resulting in a serious BLEU decline of 14.6 and 14.3.* Second, the average ΔBLEU within different CER intervals is given in Table 6. The average ΔBLEU is the sum of the BLEU decline of each sentence divided by the total # sentences within the given CER interval. It measures the impact of OCR noise at different levels on the translation quality. As shown in Table 6, sentences within [0, 1%] CER interval can be regarded as clean texts so that the translation quality remains almost unchanged. However, such sentences only account for 2.9% for *political report* and 1.4% for *scientific article*. CER of most sentences exceeds 1%. Moreover, *with the increase of CER, the BLEU decreases more severely, showing the vulnerability of the translation module to OCR noise.*

Challenges in Addressing OCR Noise. Three methods for OCR noise are experimented with. *1) Noisy text fine-tuning*: The translation module is fine-tuned with the OCR noisy text-reference bi-text to be adapted to OCR noise. *2) Post-correction*: An additional error corrector, BertChecker [7], is employed for OCR error correction. *3) Adversarial fine-tuning*: The translation module is fine-tuned with the original-adversarial sample pairs for better tolerance to OCR noise.

Table 6. Impact of OCR noise on translation within different CER intervals.

#	CER	Political Report			Scientific Article		
		# Sen	Percentage of # Sen	Avg. ΔBLEU	# Sen	Percentage of # Sen	Avg. ΔBLEU
1	[0, 1%]	117	2.9%	1.59	122	1.4%	0.45
2	(1%, 5%]	518	12.9%	8.51	490	5.7%	6.33
3	(5%, 10%]	797	19.9%	13.76	1064	12.4%	10.41
4	(10%, 20%]	645	16.1%	19.24	1340	15.6%	13.22
5	(20%, 30%]	192	4.8%	24.64	523	6.1%	12.76
6	(30%, 100%]	1738	43.4%	26.81	5060	58.8%	18.42
Total	[0, 100%]	4007	100%	19.79	8599	100%	15.33

Table 7. Performance of methods to address OCR noise. Metric is BLEU.

Methods to Address OCR Noise	OCR Noisy Source Sentence		Ground Truth Source Sentence	
	Political Report	Scientific Article	Political Report	Scientific Article
None	24.9	16.4	39.5	30.7
Noisy Text Fine-tuning	26.5 (+1.6)	17.1 (+0.7)	37.6 (–1.9)	27.7 (–3.0)
Post-correction	26.0 (+1.1)	16.8 (+0.4)	38.8 (–0.7)	29.9 (–0.8)
Adversarial Fine-tuning	27.2 (+2.3)	18.2 (+1.8)	32.9 (–6.6)	25.3 (–5.4)

As shown in Table 7, *all three methods improve BLEU, among which adversarial fine-tuning performs the best.* However, they still suffer from: 1) The deficiency which results in a large margin compared with the ground truth translation. 2) The ability degradation of clean sentence translation. Therefore, *a more balanced scheme that resists OCR noise without sacrificing the capacity to translate clean sentences, deserves further exploration.* One feasible direction is to leverage the sentence image patch for auxiliary cross-modal features.

4.2 Analysis of Layout Structure

Overall Impact. The layout degeneration problem caused by the OCR module brings two issues: 1) Truncated text fragments instead of complete sentences are transcribed. 2) The disordered arrangement of these text fragments. Such issues lead to cluttered, semantically confused source text and finally a poorly document translation. As shown in Table 8, the removal of layout analysis module leads to 11.11, 11.59 and 2.97 BLEU decline for *political report*, *scientific article*, and *paper book*, respectively, *demonstrating the indispensability of layout structure incorporation for DIT.*

Detailed Analysis. For detailed analysis, gold-standard line box and transcribed text are employed as "perfect" OCR output to shield the impact of OCR noise. The {validation set + test set} of *political report* are divided into two parts according to layout complexity: 1) Regular-layout document images with rectangular layout blocks in single column (Fig. 1 (c)(d)). 2) Irregular-layout document images with layout blocks in multiple columns or even scattered distribution (Fig. 1 (a)). Two methods are compared under this setting. *1) Base*: Layout structure is completely disregarded, *i.e.*, the transcribed text is

Table 8. Comparison between systems with/without layout analysis module.

With or Without Layout Analysis Module	Political Report				Scientific Article				Paper Book			
	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	BLEU↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
Faster R-CNN + Transformer-FT	**23.58**	**15.49**	**33.35**	**44.65**	**16.16**	**19.19**	**33.27**	**34.51**	**13.63**	**9.34**	**23.83**	**33.49**
- Faster R-CNN	12.47	9.40	22.51	33.46	4.57	6.89	19.88	19.77	10.66	8.31	22.24	31.30

Table 9. Performance comparison between method without layout and method with gold-standard layout. Evaluation metric is BLEU.

Layout Category	Regular-layout	Irregular-layout
# pages	141	139
Base	35.6	28.9
Gold-standard Layout	43.6	39.7
Δ BLEU	8.0	10.8

translated fragment by fragment and is arranged from top to bottom. *2) Gold-standard layout*: Text fragments are aggregated with the gold-standard layout box and are arranged in annotated logical order. Then, sentence segmentation and translation are performed for each text block.

As shown in Table 9, *for regular/irregular document images, the aggregation of layout structure improves BLEU by 8.0 and 10.8, respectively.* The more fine-grained BLEU improvement concerning document image proportion is shown in Fig. 4. ΔBLEU is less than 16 for 90.10% regular-layout document images. While irregular-layout document images with $\Delta\text{BLEU} \geq 16$ still account for 26.60%, indicating that *the translation improvement may be greater for irregular-layout document images.*

Challenges in Leveraging Layout. Three typical layout analysis approaches are experimented with. *1) Rule-based*: The rule-based reading flow algorithm [12], sentence segmentation and translation are applied to text fragments sequentially. *2) Projection-based*: Recursive X-Y cut [4] is employed to decompose a document image recursively into a set of rectangular layout blocks. *3) Object Detection-based*: Object detection neural network (Faster R-CNN) is used for layout block detection.

Based on Table 10, the following conclusions could be reached. 1) For regular-layout document images, the rule-based method works as well as the gold-standard layout. For irregular-layout document images, the object detection-based method performs best and reaches the same level as gold-standard layout, beating the rule-based and projection-based methods by a large margin. 2) Despite the performance superiority, the object detection-based method suffers inferior inference time - 3.5 times that of the translation module. Therefore, *for irregular-layout document images, how to make time-efficient use of layout structure, deserves further investigation.* One feasible direction is to mine the semantics and position of OCR text fragments for simultaneous layout analysis and translation instead of the current cascaded, two-stage approach.

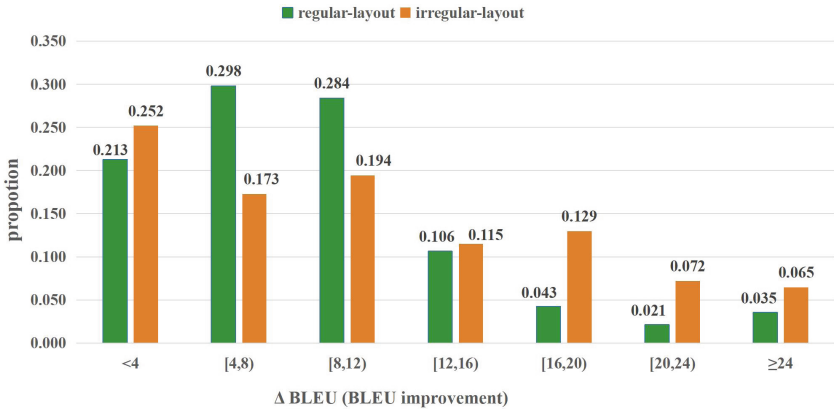

Fig. 4. Proportion of document images within different ΔBLEU intervals.

Table 10. Performance of methods of leveraging layout structure. The evaluation metric is BLEU. Avg. time delay is when processing a document image, the ratio of time consumed by layout analysis module to translation module.

Layout Analysis Methods	Translation Quality		Avg. Time Delay
	Regular-layout	Irregular-layout	
Base	35.6	28.9	–
Rule-based	42.9 (+7.3)	37.9 (+9.0)	0.03x
Projection-based	43.4 (+7.8)	38.4 (+9.5)	0.48x
Object Detection-based	43.4 (+7.8)	39.7 (+10.8)	3.5x
Gold-standard Layout	43.6	39.7	–

5 Conclusion

We developed the first document image translation dataset *DITrans* that provides three domains of document images annotated in fine granularity. In addition, benchmark experiments and detailed analysis were conducted on *DITrans* and instructive conclusions on system performance and task difficulties were drawn. With the new task, dataset and framework, we pushed a more comprehensive understanding of document images. In the future, we plan to develop models that are robust to OCR noise and make time-efficient use of the layout structure to promote DIT in both performance and efficiency.

References

1. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 296–300 (2009)
2. Cheng, Y., Tu, Z., Meng, F., Zhai, J., Liu, Y.: Towards robust neural machine translation. In: Proceedings of ACL, pp. 1756–1766 (2018)

3. Guo, H., Qin, X., Liu, J., Han, J., Liu, J., Ding, E.: Eaten: entity-aware attention for single shot visual text extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 254–259 (2019)
4. Ha, J., Haralick, R.M., Phillips, I.T.: Recursive xy cut using bounding boxes of connected components. In: Proceedings of ICDAR, pp. 952–955 (1995)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proceedings of ICCV, pp. 2980–2988 (2017)
6. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: a dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), pp. 1–6 (2019)
7. Jayanthi, S.M., Pruthi, D., Neubig, G.: NeuSpell: a neural spelling correction toolkit. In: Proceedings of EMNLP, pp. 158–164 (2020)
8. Li, M., et al.: DocBank: a benchmark dataset for document layout analysis. In: Proceedings of COLING, pp. 949–960 (2020)
9. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: Casia online and offline chinese handwriting databases. In: Proceedings of ICDAR, pp. 37–41 (2011)
10. Marti, U.V., Bunke, H.: The iam-database: an English sentence database for offline handwriting recognition. Inter. J. Document Anal. Recogn. 39–46 (2002)
11. Park, S., et al.: Cord: a consolidated receipt dataset for post-ocr parsing. In: Workshop on Document Intelligence at NeurIPS 2019 (2019)
12. Rausch, J., Martinez, O., Bissig, F., Zhang, C., Feuerriegel, S.: Docparser: hierarchical document structure parsing from renderings. In: Proceedings of AAAI, pp. 4328–4338 (2021)
13. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of NeurIPS, pp. 91–99 (2015)
14. Vaswani, A., et al.: Attention is all you need. In: Proceedings of NeurIPS, pp. 5998–6008 (2017)
15. Wang, Z., Xu, Y., Cui, L., Shang, J., Wei, F.: LayoutReader: Pre-training of text and layout for reading order detection. In: Proceedings of EMNLP, pp. 4735–4744 (2021)
16. Xu, Y., et al.: Layoutxlm: multimodal pre-training for multilingual visually-rich document understanding. ArXiv (2021)
17. Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4363–4370 (2021)
18. Zhong, X., Tang, J., Jimeno-Yepes, A.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1015–1022 (2019)