

# Query-driven Generative Network for Document Information Extraction in the Wild

Haoyu Cao<sup>\*†</sup>  
Tencent YouTu Lab  
rechycao@tencent.com

Xin Li<sup>\*</sup>  
Tencent YouTu Lab  
fujikoli@tencent.com

Jiefeng Ma<sup>‡</sup>  
University of Science and Technology  
of China  
jfma@mail.ustc.edu.cn

Deqiang Jiang  
Tencent YouTu Lab  
dqiangjiang@tencent.com

Antai Guo  
Tencent YouTu Lab  
ankerguo@tencent.com

Yiqing Hu  
Tencent YouTu Lab  
hooverhu@tencent.com

Hao Liu  
Tencent YouTu Lab  
ivanhliu@tencent.com

Yinsong Liu  
Tencent YouTu Lab  
jasonyliu@tencent.com

Bo Ren  
Tencent YouTu Lab  
timren@tencent.com

## ABSTRACT

This paper focuses on solving Document Information Extraction (DIE) in the wild problem, which is rarely explored before. In contrast to existing studies mainly tailored for document cases in known templates with predefined layouts and keys under the ideal input without OCR errors involved, we aim to build up a more practical DIE paradigm for real-world scenarios where input document images may contain unknown layouts and keys in the scenes of the problematic OCR results. To achieve this goal, we propose a novel architecture, termed Query-driven Generative Network (QGN), which is equipped with two consecutive modules, *i.e.*, Layout Context-aware Module (LCM) and Structured Generation Module (SGM). Given a document image with unseen layouts and fields, the former LCM yields the value prefix candidates serving as the query prompts for the SGM to generate the final key-value pairs even with OCR noise. To further investigate the potential of our method, we create a new large-scale dataset, named LARge-scale STructured Documents (LastDoc4000), containing 4,000 documents with 1,511 layouts and 3,500 different keys. In experiments, we demonstrate that our QGN consistently achieves the best F1-score on the new LastDoc4000 dataset by at most 30.32% absolute improvement. A more comprehensive experimental analysis and experiments on other public benchmarks also verify the effectiveness and robustness of our proposed method for the wild DIE task.

<sup>\*</sup>Both authors contributed equally to this research.

<sup>†</sup>Corresponding author.

<sup>‡</sup>Work is done during an internship at Tencent YouTu Lab.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

MM '22, October 10–14, 2022, Lisboa, Portugal.

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3547877>

## CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Applied computing** → **Document management**.

## KEYWORDS

Document Information Extraction, Visually Rich Documents, Pre-trained Models

### ACM Reference Format:

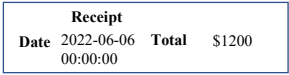
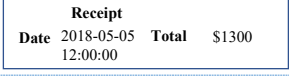
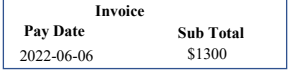
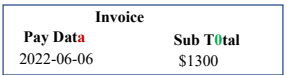
Haoyu Cao, Xin Li, Jiefeng Ma, Deqiang Jiang, Antai Guo, Yiqing Hu, Hao Liu, Yinsong Liu, and Bo Ren. 2022. Query-driven Generative Network for Document Information Extraction in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3503161.3547877>

## 1 INTRODUCTION

Document Information Extraction (DIE) aims to parse structured form consisting of key-value pairs from document images with text in the semi-structured form [1], which has become an increasingly important task in the multimedia community and plays an essential role in many downstream applications, such as business document information registration [21], verification[4, 10] and retrieval.

For the DIE application scenes, as shown in Fig. 1, we summarize them into the following three types with the order of increasing difficulty: 1) **Fixed keys and layouts without OCR noise**. Taking the receipt case for example, the keys to be extracted are usually “Date” and “Total”, while the receipts from different vendors could have different but relatively fixed key-value layouts. 2) **Unknown keys and layouts without OCR noise**. The key and layout are all agnostic to the model, which aims to verify the generalization ability in the scenario of open information extraction. 3) **Unknown keys and layouts with OCR noise**. In addition to unknown keys, OCR errors are also taken into account, which is the most practical and intractable scene we mainly investigate in this paper. We define this problem as *DIE in the wild*.

To our best knowledge, the most existing DIE methods devote to seeking solutions for the first two DIE scenes aforementioned, which can be categorized into four groups: template-based methods,

Train image	Seq	Link	Gen	QGN (ours)
				
Test 1) Fixed keys and layouts without OCR noise				
	✓	✓	✓	✓
Test 2) Unknown keys and layouts without OCR noise				
	×	✓	×	✓
Test 3) Unknown keys and layouts with OCR noise				
	×	×	×	✓

**Figure 1: Comparison of application scene between existing methods and our QGN. Existing methods are not excelling in handling the scenario with practical issues such as unseen keys, unseen layouts, and OCR errors simultaneously. “Seq”, “Link”, “Gen” are short for “sequence annotation-based methods”, “linking-based methods” and “generative-based methods” respectively. “✓” and “×” indicate whether methods support the application scenario.**

sequence annotation-based methods [19, 35, 36, 38], generative-based methods [3, 23, 34], and linking-based methods [11, 24, 31]. These works perform well in particular scenarios based on the limited assumption, *e.g.*, fixed layouts, predefined keys and without noisy OCR results. More concretely, the template-based methods employ a layout matching strategy assuming all documents in fixed layout and generated with the same template. In a similar sense, all keys should be predefined and known in advance in sequence annotation-based and most generative-based methods, while linking-based methods work on the condition of totally correct OCR results organized in reading order. However, it is difficult to obtain the correct reading order in the real-world application because of the enormous variety of layouts and keys existing in the real world. Therefore, the DIE in the wild is still a rarely explored and challenging task.

To solve this practical problem, we propose a novel model following the query-then-generation pipeline, termed Query-driven Generative Network (QGN), which can adequately combine the merits of existing linking-based methods and generative-based solutions. Compared with linking-based methods that 1) first predict classes of manually ordered entities and then 2) link them, QGN only predicts the **value prefixes** of entities at the first stage, which provides simpler but more robust **query prompts** [9] for guiding the next generation step without strict reliance on OCR. Given the query prompts, QGN follows a “divide and conquer” paradigm: instead of generating all key-value information in one action, local key-value generations are conducted by referring to each prompt. The generated local results are aggregated into the final output afterward, which could alleviate the accumulated errors incurred by generative-based methods. Consequently, the difficulty of each sub-task is properly reduced in our proposed QGN, which is able to achieve surprisingly better performance than existing methods.

Moreover, we also observe that our model can well handle the OCR errors in more practical scenes.

Additionally, to further explore the potential of our method under various scenarios of layouts and fields, we establish a new large-scale dataset, which contains 4,000 documents, 1,511 different types of layouts, and more than 3,500 various keys, denoted as LArge-scale STructured Documents dataset (LastDoc4000). Our main contributions can be summarized as follows:

- 1) This paper defines a more challenging Document Information Extraction (DIE) task, termed DIE in the wild, where input document images may contain unknown keys and layouts with problematic OCR involved, which has been rarely explored in previous works.
- 2) We propose a novel query-based generative architecture, QGN, tailored for DIE in the wild task. Especially, thanks to the query-extraction and pair-generation mechanisms, QGN no longer relies on the faultless OCR results as well as the limited predefined keys, while required by other methods.
- 3) Our proposed QGN can not only achieve comparable state-of-the-art performance on public benchmarks, but also on a more challenging dataset we newly collected, which demonstrates the effectiveness of the proposed QGN and further confirms its applicability under real scenarios.

## 2 RELATED WORK

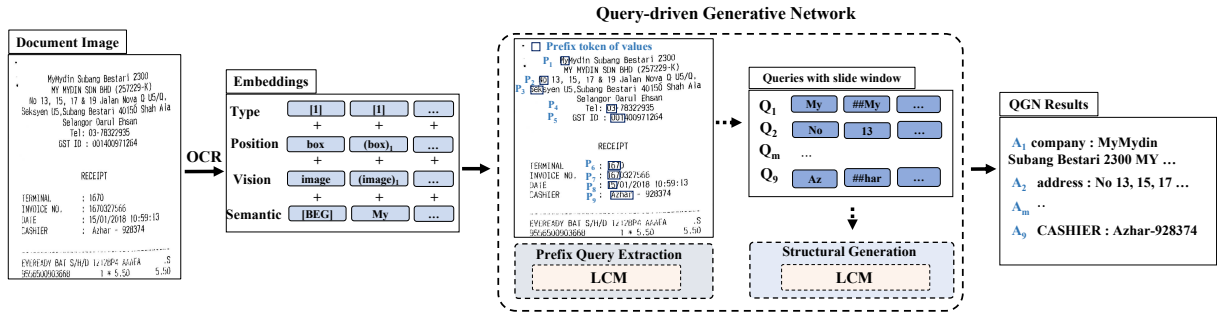
Before deep learning became widespread in DIE, traditional methods of this task [8, 28] heavily relied on the predefined template rules, which could not generalize across different document layouts. Recently, significant improvements have been achieved both in performance and robustness via deep-learning-based methods. These approaches can be mainly categorized into three groups: sequence annotation-based methods, generation-based methods, and linking-based methods.

### 2.1 Sequence annotation-based methods

These methods aim to classify each document token into predefined categories, similar to the named entity recognition paradigm. In early, Sage *et al.* [27] take semantics and layout information into account and uses recurrent neural networks to extract entities of interest. Recently, considerable progress has been achieved by multimodal-based models [17–19, 35–37] which encode visual, semantic, and layout modalities into the Bert-based [6] model. However, content serialization is still a heavy dependency for these methods, indicating that all texts should be prepared in correct reading order first, which is practically difficult due to layouts’ diversity in the real world.

### 2.2 Generative-based methods

In order to efficiently employ weak supervision with document-level annotations instead of costly word-level annotations, Powalski *et al.* [23] utilize a T5 [25] based encoder-decoder Transformer [32] for end-to-end DIE training, and Sage *et al.* [26] adopt a PGN-based [29] decoder separately. Moreover, Wang *et al.* [34] train a flexible decoder with the generation and tagging objectives simultaneously optimized. These generative-based methods comply with



**Figure 2: The architecture of our proposed method. LCM-based backbones in the extraction and generation stage are designed in the same structure with parameters sharing. “Type” is a flag embedding to distinguish the two stages, semantic embeddings with “##” indicate the word-piece inputs.  $P_i$ ,  $Q_i$ , and  $A_i$  respectively represent prefix of values, corresponding queries, and generated results. Best viewed in color.**

the seq2seq pattern, which can deal with some OCR misrecognition. However, they still have an implicational dependency on the correct sequential order through 1D position embedding and could not cover unseen keys due to limited and predefined categories.

### 2.3 Linking-based methods

These methods mainly focus on scenes of unlimited keys, including the graph-based method and key-value linking method. For the former, MatchVIE [31], SPADE [13] and BROS [11] propose a graph-based decoder to extract key contexts from identified connectivity between text blocks. For the others [36], a key-value entity recognition model along with the classification-based relation extraction is employed for the multilingual DIE task. Note that, part of the multimodal pre-training models mentioned in the sequence annotation-based methods can also be applied to this kind of method due to the model architecture is similar as LayoutLM. These methods perform well on undefined keys of idealized scenes but could be seriously influenced by the OCR errors.

## 3 METHOD

### 3.1 Overall Architecture

The overview of the proposed Query-driven Generative Network (QGN) is shown in Fig. 2, which consists of 1) Prefix Query Extraction (Section 3.3), 2) Structural Generation Module (SGM, Section 3.4), equipped with 3) Layout Context-aware Module (LCM, Section 3.2). Firstly, multi-modality embeddings, namely vision, text, position and type embeddings are extracted similarly with LayoutLM [35]. Afterward, the specific query prefix of each value entity is picked up using the window mechanism, based on which the query vectors are aggregated with context-sensitive modalities. Furthermore, key-value pairs are generated word-by-word at the generation stage, with each query vector acting as the prompt. Note that a generate-with-copy mechanism is employed to decide whether to generate new words or copy existing ones from source texts.

### 3.2 Layout Context-aware Module

LCM takes multi-modality embeddings of each token as input and outputs the corresponding representations with fixed dimensions, following the architecture of the Transformer [32]. In previous

Transformer-based works [13, 37], relative position information is employed to model the local invariance of document layout. However, there exist two limitations in the relative position bias: 1) the positional embedding is unable to capture contextual information in complex scenes (*e.g.*, the key and matched value are seriously misaligned), and 2) utilizing the inductive bias of all tokens to compute attention matrix may involve noisy tokens and weaken the local perception, leading to potential degradation of performance. To address these issues, we propose a new scheme named LCM, stacked by Layout Context-aware Block (LCB), which can aggregate the *Vision*, *Semantic* and *Position* information together for better capturing effective contextual layout bias, as shown in Fig. 3. Concretely, we consider a measurement  $R_{ij}^l$ , which measures various relative information  $l \in \{V, S, P\}$  between  $i$ -th token and  $j$ -th token:

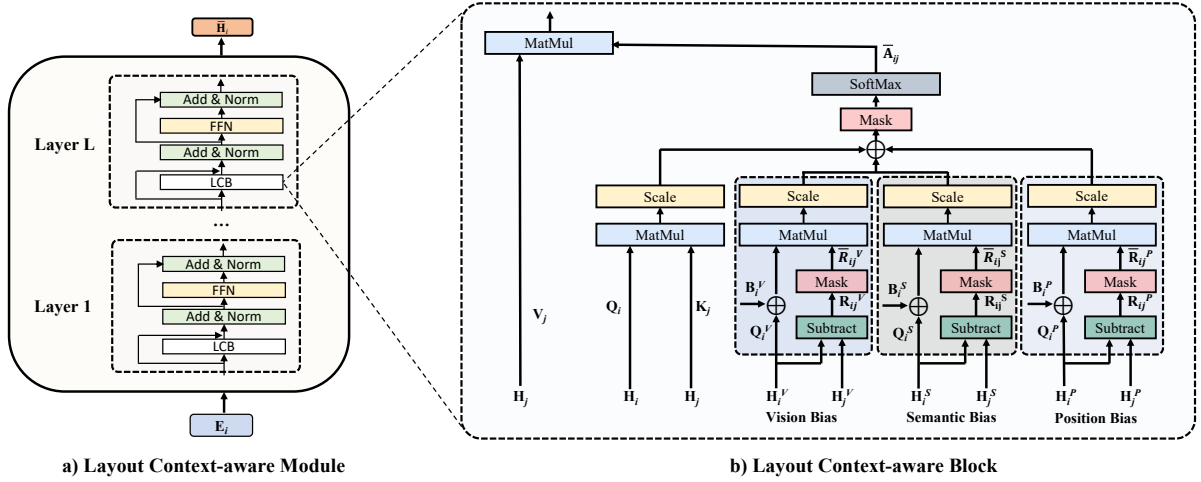
$$R_{ij}^l = H_i^l - H_j^l, \quad (1)$$

where  $H_i^l \in \mathbb{R}^{1 \times d_h}$  and  $H_j^l \in \mathbb{R}^{1 \times d_h}$  denote the feature of  $i$ -th token and  $j$ -th token,  $i, j \in N$ ,  $N$  is the amount of input tokens,  $d_h$  means the dimension, the subtract function is employed to get the feature differences measured by the Euclidean distance between two tokens. The attention weights between each key  $K$  and query vector  $Q$  are calculated by:

$$A_{ij} = Q_i^T K_j + \sum_l (Q_i^{T(l)} + B_i^l)(R_{ij}^l), \quad (2)$$

where  $Q_i$ ,  $Q_i^T$  are the query vector and bias query vector of the  $i$ -th input token separately.  $K_j$  is the key vector of the  $j$ -th input token,  $B_i^l$  is a bias vector.

Instead of involving all the query-bias pairs into the attention matrix as the inductive bias of layout, we only focus on the similar tokens for each query and mask the other irrelevant ones. To be specific, for the  $i$ -th element, we calculate the Euclidean distance against all other elements and get its  $k$ -nearest neighbors, discarding the rest ones, then the masked matrix  $M^l$  is built with the mask ratio of  $\Delta$  correspondingly. The extended query-key attention



**Figure 3: An illustration of our proposed Layout Context-aware Module (LCM) stacked with  $L$  layers of Layout Context-aware Blocks (LCB).  $\bar{H}_i$  indicates the multi-modal feature output by LCM with the  $i$ -th input token embedding  $E_i$ .  $H_i^V$ ,  $H_i^S$  and  $H_i^P$  represent vision, semantic and position embedding respectively. Best viewed in color.**

product matrix  $\bar{A}_{ij}$  can be noted:

$$\bar{R}_{ij}^l = R_{ij}^l M^l, \quad (3)$$

$$\bar{A}_{ij} = Q_i^T K_j + \sum_l (Q_i^{(l)} + B_i^l) (\bar{R}_{ij}^l), \quad (4)$$

where  $\bar{R}_{ij}^l$  represents the masked encoding.

### 3.3 Prefix Query Extraction

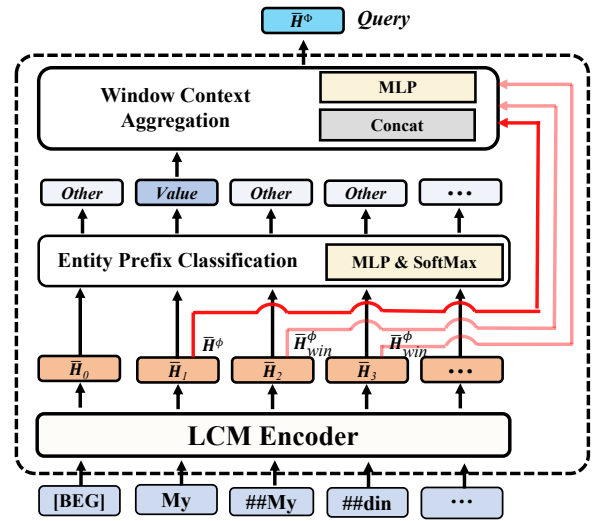
In previous works [35–37], entity extraction (*i.e.*, key and value) is often the first step of token-level sequence tagging. Unfortunately, the extraction usually suffers from entity boundary ambiguity problems caused by noisy text recognition and wrong reading order, which leads to performance degradation.

To attack the problem, we propose a novel Prefix Query Extraction submodule to relieve the reliance on entity integrity, which only extracts the entity prefix by combining entity extraction and window sampling. As shown in Fig. 4, a linear mapping equipped with fully-connected layers and softmax layer is applied to classify the hidden state of each token produced by LCM, which determines whether it is the prefix token of the value or not. Given the  $i$ -th token's feature  $\bar{H}_i \in \mathbb{R}^{1 \times d_h}$ , the function of entity prefix classification can be defined as:

$$F_i^\phi = \text{SoftMax}(\bar{H}_i W_\phi), \quad (5)$$

where  $F_i^\phi$  demonstrates the confidence of classification type,  $\phi \in \{Value, Other\}$ ,  $W_\phi \in \mathbb{R}^{d_h \times d_\phi}$  is a trainable projection matrix,  $d_\phi = 2$ . Then, prefix tokens  $\bar{H}^\phi \in \mathbb{R}^{N \times d_h}$  of *Value* serve as queries with contextual information after classification, where  $N$  is the amount of selected queries.

In order to enrich the information of the selected token, a local window mechanism is adopted to fuse adjacent tokens' features together. Here the window is defined as continuous sequential

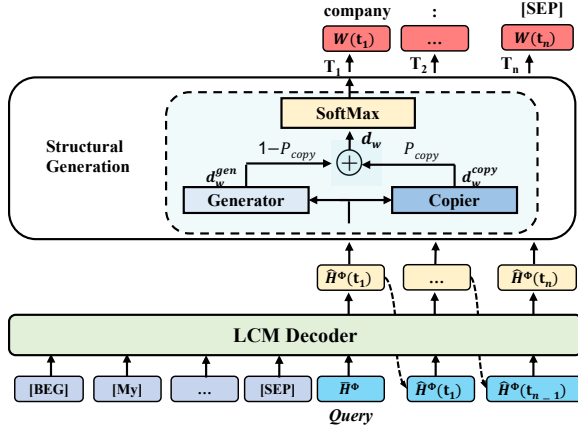


**Figure 4: The diagram of Prefix Query Extraction module. The query *Value* with light blue color is output by entity prefix classification, which is aggregated with window context to generate the final query vector  $\bar{H}^\Phi$ . Best viewed in color.**

tokens after the prefix query:

$$\bar{H}^\Phi = \text{Concat}(\bar{H}^\phi, \{\bar{H}_{win}^\phi\}) W_\Phi, \quad (6)$$

where  $\{\bar{H}_{win}^\phi\} \in \mathbb{R}^{N \times d_h \times \Lambda}$  is the set of adjacent tokens in the local window,  $\Lambda$  is the window size,  $W_\Phi \in \mathbb{R}^{(\Lambda+1) \times d_h \times d_h}$  is the linear projection of window aggregation,  $\bar{H}^\Phi \in \mathbb{R}^{1 \times d_h}$  denotes the output states of the query with window features. Compared with the previous methods relying on entire entity tagging, the prefix classification scheme is simple yet effective to fully utilize the contextual layout information.



**Figure 5: The architecture of Structural Generation stage.** “ $T_n$ ” denotes the  $n$ -th timestep generation within one query process. After each timestep, the hidden state  $\hat{H}^\Phi(t)$  produced by LCM is appended to the input sequence for the next decoding step. The query process comes to an end with a “[SEP]” being predicted. Best viewed in color.

### 3.4 Structural Generation

Given the aggregated queries  $Value$ , the results of key-value pairs will be generated word-by-word in the subsequent Structural Generation process, which is mainly implemented by an LCM decoder and a copy-and-generative block, as depicted in Fig. 5. For the input, the query vector  $\hat{H}^\Phi$  is appended to the end of the original sequence of length  $N$ , and the position embedding of decoded tokens are built with the 2D coordinate as the LCB does.

Taking the query  $\hat{H}^\Phi$  as input, the LCM outputs the corresponding contextual state  $\hat{H}^\Phi(t)$  with the hidden size of  $d_h$ , where  $t$  represents the  $t$ -th timestep of decoding. Furthermore, the copy mechanism is inherited from an effective text summarization method, PGN [29], to improve the performance. Specifically, for each timestep  $t$ , a generation probability distribution  $d_w^{gen}(t)$  of each word is generated by the generator. Meanwhile, another copy probability distribution  $d_w^{copy}(t)$  is also generated by the copier which indicates whether the candidate word is in the source text. The final distribution is calculated as:

$$d_w(t) = (1 - P_{copy}(t))d_w^{gen}(t) + P_{copy}(t)d_w^{copy}(t), \quad (7)$$

where  $P_{copy}(t) \in [0, 1]$ , used as a soft switch to choose a word between generation or copied from the input sequence. For each timestep,  $\hat{H}^\Phi(t)$  output by LCM decoder is appended to the sequence, which is regarded as a new context vector to guide the output for the next timestep:

$$\hat{H}_{input}^\Phi(t+1) = \hat{H}_{output}^\Phi(t). \quad (8)$$

When the field terminator [SEP] is generated, the entire query process comes to an end, likewise the key-value pair generation. Finally, the generated results from all queries compose the structured information result.

## 3.5 Training Strategy

**3.5.1 Design of loss function.** The proposed QGN is trained with the multiple optimization tasks of entity prefix classification and structured generation in an end-to-end way. The global optimization can be defined as:

$$\mathcal{L} = \lambda_1 \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{class} + \lambda_2 \frac{1}{T} \sum_{t=1}^T \mathcal{L}_{gen}, \quad (9)$$

where  $\mathcal{L}_{class}$  and  $\mathcal{L}_{gen}$  represent the cross-entropy loss of entity prefix classification and information generation loss respectively, which are combined by weight parameters  $\lambda_1$  and  $\lambda_2$ .  $N$  is the amount of input tokens,  $T$  is the number of timesteps.

**3.5.2 Pre-training.** To improve the performance of QGN, we use large-scale data for pre-training. On one hand, we sample 6 million English documents from the IIT-CDIP [16] dataset partially following LayoutXLM [36]. On the other hand, Chinese and English documents obtained from the web are also employed, consisting of 37 million metadata after cleaning. Hence, the total amount of documents for pre-training is about 43 million.

In the pre-training stage, we simultaneously use three self-supervised tasks derived from UniLM [7], including Unidirectional LM, Bidirectional LM, and Sequence-to-Sequence LM. Specifically, during one training batch, the bidirectional LM objective occupies 1/3 of the time, and the sequence-to-sequence LM objective also occupies 1/3 of the time, while both left-to-right and right-to-left LM objectives occupy 1/6 of the time respectively. The token masking probability is 15%, among which 80% are replaced by token [MASK], 10% by a random token, and the rest are unchanged.

**3.5.3 Data augmentation.** To achieve further robustness in the wild scenes, we introduce three data augmentation techniques to the proposed QGN: semantic, spatial, and visual augmentation, respectively.

**Semantic augmentation.** In order to simulate OCR recognition errors and make the structural generation compatible with such cases, we randomly replace 10% of the input texts with other characters in the vocabulary while keeping the target structured results unchanged. Note that no digits will be replaced due to the lack of contextual information.

**Spatial augmentation.** Considering the high diversity of layouts, we introduce a geometric transformation to enrich the layout distribution. In detail, slight jittering is employed to dynamically generate the offset of bounding boxes, with the ratio range in  $[0, 0.25]$  of the average box height. Besides, a global scaling is performed on the coordinates of boxes with a random coefficient in the range  $[0.65, 1.25]$ .

**Visual augmentation.** In the wild scenes, the image capturing device usually suffers from various perspective distortions and illumination problems, which lead to undesirable image flaws. In order to address this issue, affine transformation and image enhancement are performed to simulate the visual content variances of input images.



**Table 1: Comparison of different settings.** “Already seen” and “Unseen” indicate whether the templates have been seen in the training set, “GT” and “Realistic” indicate using the ground truth or realistic OCR as input.

Settings	Layout and keys		OCR Input	
	Already Seen	Unseen	GT	Realistic
Setup-A	✓		✓	
Setup-B	✓			✓
Setup-C		✓	✓	
Setup-D		✓		✓

## 4 EXPERIMENTS

### 4.1 Experimental Settings

Existing DIE solutions usually exploit normalized inputs without OCR errors and unknown keys. In this task, we conduct the following experiments 1) Setup-A: using the ground truth as the input, 2) Setup-B: using realistic OCR results as input, 3) Setup-C: based on Setup-A and with unseen layouts and unseen keys appear, 4) Setup-D: based on Setup-C, and using realistic OCR results as input, as shown in Tab. 1. We employ SBDNet [20] with CRNN [30] as OCR engines for OCR input.

### 4.2 Implement Details

**4.2.1 Pre-training Implementation details.** The proposed QGN stack 12 layers of LCB with 8 attention heads, 768 for hidden size. Distinct from the original Transformer, the model is partially initialized from the released LayoutXML<sub>Base</sub> with the newly introduced modules initialized by random weights. Besides, AdamW [15] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  is used for optimization. The learning rate is  $5e-5$ , with linear warm-up over the first 40,000 steps and linear decay, and the weight decay is 0.01. The maximum length of the input sequence is 512. We train the model with a total batch size of 1,024 for 5 epochs by 16 Nvidia Telsa A100 80GB GPU.

**4.2.2 Fine-tuning Implementation details.** The hyper-parameter of maximum sequence length is 1536 and the loss weight parameters  $\lambda_1 = 0.3$ ,  $\lambda_2 = 0.7$  for all downstream tasks. During training, the model is optimized by the Adam with the initial learning rate of  $5e-5$ , and a linear decay learning rate schedule. Note that all experiments are conducted under the same environment, in which the model is first pre-trained with a total batch size of 1,024 for 5 epochs and then fine-tuned with batch size 48 on various datasets for 100 epochs.

### 4.3 Comparison with State-of-the-arts

**4.3.1 Fine-tuning Datasets.** We conduct extensive experiments on five public datasets of Document Information Extraction (DIE) and the proposed LastDoc4000 dataset, as summarized in Tab. 2.

**SROIE** [12] consists of 973 scanned receipts in English, each of which is annotated with GT text, corresponding bounding boxes, and entity type (*i.e.*, Company, Address, Date, and Total).

**CORD** [22] is also a receipt dataset collected from various sources, including shops, restaurants, *etc.*, which contains 800 images for the training set, 100 for the validation set, and 100 for the testing set.

**Table 2: Statistics of various datasets in our experiments.**

Dataset	Train+Dev	Test	Entity
SROIE	626	347	4
CORD	900	100	30
EPHOIE	1183	311	10
FUNSD-R	149	50	-
XFUND-ZH-R	149	50	-
LastDoc4000	2687	1313	3519

**Table 3: Comparison results of DIE on SROIE, CORD, EPHOIE, FUNSD-R, and XFUND-ZH-R datasets under Setup-A and Setup-B settings. F1 means F1-score. The subscript “B” stands for “Base”, “L” stands for “Large”.**

SROIE		
Method	Setup-A	Setup-B
	F1(%)	F1(%)
BERT <sub>B</sub> [6]	90.99	73.27
UniLMv2 <sub>B</sub> [2]	94.59	-
LayoutXML <sub>B</sub> [36]	94.80	79.40
BERT <sub>L</sub> [6]	92.00	75.15
UniLMv2 <sub>L</sub> [2]	94.88	-
LayoutLM <sub>L</sub> [35]	95.24	79.31
LayoutLMv2 <sub>L</sub> [37]	97.81	-
BROS [11]	95.48	-
PICK [38]	96.12	-
VIES [33]	96.12	-
TCPN [34]	96.54	91.93
MatchVIE [31]	96.57	-
StrucTexT [19]	96.88	-
<b>QGN ours</b>	<b>97.90</b>	<b>92.27</b>
CORD		
BERT <sub>B</sub> [6]	89.68	61.72
UniLMv2 <sub>B</sub> [2]	91.98	-
LayoutXML <sub>B</sub> [36]	94.84	67.36
BERT <sub>L</sub> [6]	90.25	64.38
UniLMv2 <sub>L</sub> [2]	92.05	-
LayoutLM <sub>L</sub> [35]	94.93	69.36
LayoutLMv2 <sub>L</sub> [37]	96.01	-
BROS <sub>L</sub> [11]	<b>97.28</b>	-
<b>QGN ours</b>	96.84	<b>83.03</b>
EPHOIE		
LayoutXML <sub>B</sub> [36]	97.69	74.69
VIES [33]	95.23	83.81
TCPN [34]	98.06	86.19
MatchVIE [31]	96.87	-
<b>QGN ours</b>	<b>98.49</b>	<b>89.25</b>
FUNSD-R		
LayoutXML <sub>B</sub> [36]	52.05	30.11
<b>QGN ours</b>	<b>54.82</b>	<b>39.79</b>
XFUND-ZH-R		
LayoutXML <sub>B</sub> [36]	64.47	48.45
<b>QGN ours</b>	<b>65.45</b>	<b>62.21</b>

**Table 4: F1(%) scores of different DIE methods on the LastDoc4000 under various setups. “\*” indicates the results reproduced by the official provided code and pre-trained models.**

Method	LastDoc4000			
	Setup-A	Setup-B	Setup-C	Setup-D
	F1(%)	F1(%)	F1(%)	F1(%)
InfoXMLM <sub>B</sub> <sup>*</sup> [5]	60.69	38.30	43.00	28.95
LayoutXMLM <sub>B</sub> <sup>*</sup> [36]	82.84	58.61	64.13	50.92
<b>QGN ours</b>	<b>89.16</b>	<b>83.86</b>	<b>86.55</b>	<b>81.24</b>

The dataset defines 30 entity categories based on GT annotations of each receipt.

**EPHOIE** [33] is composed of 1,494 examination paper heads collected from Chinese schools’ examinations, which are annotated with ten types of entities (*i.e.*, School, Class, Name).

**FUNSD** [14] contains 199 noisy scanned documents as well as corresponding annotations. Due to FUNSD only annotate entities and link relations, we relabel it with key-value pairs formats, termed FUNSD-R.

**XFUND** [36] is a multilingual form understanding dataset, same as FUNSD, we relabel the Chinese subset with key-value pairs formats, termed XFUND-ZH-R, which containing 199 scanned documents. **LastDoc4000** is a new large-scale structured documents dataset collected from the wild scenes, which contains 4,000 images with 1,511 layouts and more than 3,500 different keys. Details of the LastDoc4000 dataset are given in the appendix.

The F1-score is applied to measure the entity level accuracy of DIE task, which is calculated with the whole key-value pair strings.

**4.3.2 Results on SROIE, CORD, EPHOIE, FUNSD-R and XFUND-ZH-R datasets.** As shown in Tab. 3, our proposed approach achieves comparable state-of-the-arts on different datasets under both Setup-A and Setup-B settings, which demonstrates its superior performance. Structured results under the setting of Setup-B are visualized in Fig. 6 (a)-(d), which reveals that QGN is robust against OCR errors and support error correction. For the Setup-A metrics, we extract from the corresponding paper, and for Setup-B, we fine-tuned the corresponding models reference to the official released code and pre-training models, *i.e.* Bert[6], LayoutLM[35], LayoutXMLM[36]. Specifically, baseline models are fine-tuned with the sequence annotation-based mode on the SROIE, CORD, EPHOIE dataset, and linking-based mode on other datasets.

**4.3.3 Results on LastDoc4000 dataset.** The results on LastDoc4000 are summarized in Tab. 4. Compared with the strong baseline LayoutXMLM, our method improves 6.32% and 25.25% under Setup-A and Setup-B. The gain is similar to other datasets. In Setup-C and Setup-D, our QGN achieves at least 22.42% improvement over other methods, as shown in Fig. 6 (e). Compared with “unseen” and “already seen” results, QGN shows minor performance drops while baseline models suffer from huge performance degradations, which indicates that QGN has a great advantage to handle zero-shot scenarios, as shown in Fig. 7.

	Noisy OCR input	Results of LayoutXMLM	Results of our QGN
a)			
b)			
c)			
d)			
e)			

**Figure 6: Visualization results of LayoutXMLM and QGN on various benchmarks under setup-B. Key entities are highlighted in green, while value entities are in yellow. The images in rows of a), b), c), d) and e) are samples from SROIE, CORD, EPHOIE, XFUND-ZH-R, LastDoc4000, respectively. The second column presents the results of LayoutXMLM, with the error ones highlighted in red. By contrast, our QGN is able to yield the correct results given in the third column, which are marked in green color. Our QGN shows better robustness in the noisy scenes compared with LayoutXMLM.**

	GT Input	Results of LayoutXMLM	Results of our QGN
a)			
b)			
c)			

**Figure 7: Visualization results of LayoutXMLM and QGN on the LastDoc4000 dataset under setup-C.**

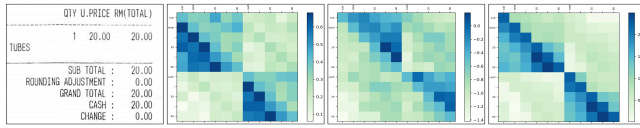
#### 4.4 Ablation Study and Further Discussion

To analyze the effectiveness of different submodules in our designed network, we perform the following extensive experiments on the LastDoc4000 dataset under Setup-D. Effect of augmentation strategy is given in the appendix.

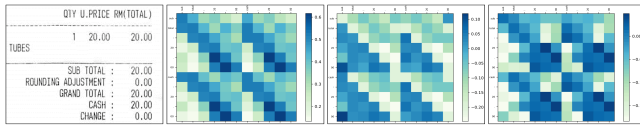
**4.4.1 Effect of LCM.** We validate the effectiveness of proposed encodings in LCM. The results are summarized in Tab. 5, from which we can see that the position encoding increases F1-score by 1.70%. Moreover, additional vision and semantic encoding contribute 1.15% and 2.38% gain, separately. To explore the effectiveness of layout encodings, we also visualize the attention maps from the first layer of LCM shown in Fig. 8. The attention weights of position tend to be local, indicating that the relative positional encoding enhances localized contextual representation. Moreover, the attention weights of visual and semantic encodings are inclined to emphasize the

**Table 5: Ablation study of encodings and mask in LCM on LastDoc4000 dataset.** “P”, “V” and “S” are short for “Position”, “Vision” and “Semantic” encoding individually. “M” and “ $\Delta$ ” are “Mask” and the mask ratio. “w/” and “w/o” stand for “with” and “without” respectively.

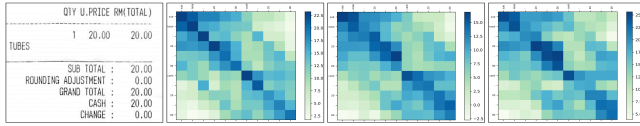
Method	P	V	S	M	$\Delta$	Setup-D
						F1(%)
QGN <sub>w/o bias</sub>	✗	✗	✗	✗	-	74.54
QGN <sub>w/ P</sub>	✓	✗	✗	✗	-	76.24
QGN <sub>w/ P&amp;V</sub>	✓	✓	✗	✗	-	77.39
QGN <sub>w/ P&amp;V&amp;S</sub>	✓	✓	✓	✗	-	79.77
QGN <sub>w/ P&amp;V&amp;S&amp;M</sub>	✓	✓	✓	✓	0.3	<b>81.24</b>



(a) Distribution of position encoding heat map.



(b) Distribution of vision encoding heat map.



(c) Distribution of semantic encoding heat map.

**Figure 8: Visualization of the heat-maps generated by Layout Context-aware block.** The colored blocks quantify how much attention the  $i$ -th token pays to the  $j$ -th token in various decodings.

words with similar features, which indicates that the model learns more inductive bias from variable layout encodings. We argue that these different encodings exhibit complementary attention scores emphasis, which boosts the performance of the subsequent DIE task. To validate the effect of hyper-parameters of QGN, we separate 10% of the training set for validation. For the mask ratio on LCB, we start with 0, and in increment of 0.1, the performance keeps improving until  $\Delta = 0.3$ , which indicates the effectiveness of the proposed mask mechanism to denoise the attention of irrelevant. Specifically, we observe that the F1-score decreases when the mask ratio is over 0.5, the possible reason is that a higher mask ratio on layout encodings tends to shrink the receptive field. We set the hyper-parameter to 0.3 for all the tasks according to the above observation.

**4.4.2 Effect of query method.** As shown in Tab. 6, we also investigate the effectiveness of the entity prefix query mechanism. To perform the comparison, an additional sequence tagger is trained

**Table 6: Ablation study of different query schemes on the LastDoc4000 dataset.** “WhK” and “WhV” mean the whole *Key* and *Value* entity extracted by sequence tagging, while “PrK” and “PrV” stand for the prefix classification of *Key* and *Value*, respectively. “W” represents the window mechanism. “ $\Lambda$ ” denotes the window size.

Method	WhK	WhV	PrK	PrV	W	$\Lambda$	Setup-D
							F1(%)
QGN <sub>w/ WhK</sub>	✓	✗	✗	✗	✗	-	61.18
QGN <sub>w/ WhV</sub>	✗	✓	✗	✗	✗	-	69.35
QGN <sub>w/ PrK</sub>	✗	✗	✓	✗	✗	-	69.59
QGN <sub>w/ PrV</sub>	✗	✗	✗	✓	✗	-	78.66
QGN <sub>w/ PrV&amp;W</sub>	✗	✗	✗	✓	✓	4	<b>81.24</b>

**Table 7: Ablation study of the copy-generative module on the LastDoc4000 dataset.**

Method	Generative	Copy	Setup-D
			F1(%)
QGN <sub>w/ Copy</sub>	✗	✓	65.26
QGN <sub>w/ Generative</sub>	✓	✗	79.77
QGN <sub>w/ Generative&amp;Copy</sub>	✓	✓	<b>81.24</b>

to extract the entire entities. Compared with taking the whole entity as input, our prefix-based method can significantly increase F1-score by at least 8.41%, which shows its better robustness and flexibility in the noisy scenes. Additionally, applying the prefix of *Value* for structural generation is able to increase F1-score by 9.07% as clearly demonstrated by QGN<sub>w/ PrK</sub> and QGN<sub>w/ PrV</sub>. For the hyper-parameter of window size  $\Lambda$ , we set  $\Lambda = 4$  via a similar validation process.

**4.4.3 Effect of copy-generative module.** We also perform experiments to evaluate the copier and generator in the structural generative stage. The results are reported in Tab. 7, from which we can see that the generative branch plays a dominant role, which may be attributed to the Sequence-to-Sequence LM of the pre-training phase. We observe some failure cases in Fig. 6 d), which is due to the duplicated and error results in the generation stage, but it can be alleviated by the copy branch to some extent. Specifically, the copy mechanism brings 1.47% improvement.

## 5 CONCLUSION AND LIMITATION

This paper proposes QGN, a query-driven generative network for DIE in the wild task, especially for unseen layouts and keys, as well as OCR errors. Extensive experiments show that QGN outperforms the state-of-the-art methods in public DIE datasets. Additionally, we introduce a large-scale DIE dataset called LastDoc4000, a diverse benchmark for DIE, and wish it could inspire researches on DIE in the future. Generative paradigm will inevitably encounter the problem of faithfulness and factuality, we will introduce more prior knowledge as guidance to generate the structured results with higher quality and further adapt our model to more expansive application scenes in the future work.



## REFERENCES

- [1] Olivier Augereau, Nicholas Journet, and Jean-Philippe Domenger. 2013. Semi-structured document image matching and recognition. In *Document Recognition and Retrieval XX, part of the IS&T-SPIE Electronic Imaging Symposium, Burlingame, California, USA, February 5-7, 2013, Proceedings (SPIE Proceedings, Vol. 8658)*, Richard Zanibbi and Bertrand Couasnon (Eds.). SPIE, 865804. <https://doi.org/10.1117/12.2003911>
- [2] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning, ICM 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 642–652. <http://proceedings.mlr.press/v119/bao20a.html>
- [3] Haoyu Cao, Jiefeng Ma, Antai Guo, Yiqing Hu, Hao Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. GMN: Generative Multi-modal Network for Practical Document Information Extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3768–3778. <https://aclanthology.org/2022.naacl-main.276>
- [4] Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. 2020. One-shot Text Field labeling using Attention and Belief Propagation for Structure Information Extraction. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 340–348. <https://doi.org/10.1145/3394171.3413511>
- [5] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 3576–3588. <https://doi.org/10.18653/v1/2021.naacl-main.280>
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [7] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 13042–13054. <https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html>
- [8] Daniel Esser, Daniel Schuster, Klemens Muthmann, Michael Berger, and Alexander Schill. 2012. Automatic indexing of scanned documents: a layout-based approach. In *Document Recognition and Retrieval XIX, part of the IS&T-SPIE Electronic Imaging Symposium, Burlingame, California, USA, January 25-26, 2012, Proceedings (SPIE Proceedings, Vol. 8297)*, Christian Viard-Gaudin and Richard Zanibbi (Eds.). SPIE, 82970H. <https://doi.org/10.1117/12.908542>
- [9] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3816–3830. <https://doi.org/10.18653/v1/2021.acl-long.295>
- [10] Ahmed Hamdi, Elodie Carel, Aurélie Joseph, Mickaël Coustaty, and Antoine Doucet. 2021. Information Extraction from Invoices. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12822)*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Springer, 699–714. [https://doi.org/10.1007/978-3-030-86331-9\\_45](https://doi.org/10.1007/978-3-030-86331-9_45)
- [11] Teakgyu Hong, Donghyun Kim, Mingji Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2021. BROS: A Layout-Aware Pre-trained Language Model for Understanding Documents. *CoRR* abs/2108.04539 (2021). arXiv:2108.04539 <https://arxiv.org/abs/2108.04539>
- [12] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019, IEEE*, 1516–1520. <https://doi.org/10.1109/ICDAR.2019.00244>
- [13] Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial Dependency Parsing for Semi-Structured Document Information Extraction. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 330–343. <https://doi.org/10.18653/v1/2021.findings-acl.28>
- [14] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. In *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019, IEEE*, 1–6. <https://doi.org/10.1109/ICDARW.2019.10029>
- [15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1412.6980>
- [16] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a Test Collection for Complex Document Information Processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, Washington, USA) (SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 665–666. <https://doi.org/10.1145/1148170.1148307>
- [17] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021. StructuralLM: Structural Pre-training for Form Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 6309–6318. <https://doi.org/10.18653/v1/2021.acl-long.493>
- [18] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Self-Doc: Self-Supervised Document Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 5652–5660. [https://openaccess.thecvf.com/content/CVPR2021/html/Li\\_SelfDoc\\_Self-Supervised\\_Document\\_Representation\\_Learning\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Li_SelfDoc_Self-Supervised_Document_Representation_Learning_CVPR_2021_paper.html)
- [19] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Erhui Ding. 2021. StructText: Structured Text Understanding with Multi-Modal Transformers. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueqing Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metzke, and Balakrishnan Prabhakaran (Eds.). ACM, 1912–1920. <https://doi.org/10.1145/3474085.3475345>
- [20] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. 2019. Omnidirectional Scene Text Detection with Sequential-free Box Discretization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 3052–3058. <https://doi.org/10.24963/ijcai.2019/423>
- [21] Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation Learning for Information Extraction from Form-like Documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 6495–6504. <https://doi.org/10.18653/v1/2020.acl-main.580>
- [22] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.
- [23] Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. 2021. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12822)*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida (Eds.). Springer, 732–747. [https://doi.org/10.1007/978-3-030-86331-9\\_47](https://doi.org/10.1007/978-3-030-86331-9_47)
- [24] Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2019. GraphIE: A Graph-Based Framework for Information Extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 751–761. <https://doi.org/10.18653/v1/n19-1082>
- [25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. <http://jmlr.org/papers/v21/20-074.html>
- [26] Clément Sage, Alex Aussem, Véronique Eglin, Haytham Elghazel, and Jérémy Espinas. 2020. End-to-End Extraction of Structured Information from Business

- Documents with Pointer-Generator Networks. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP@EMNLP 2020, Online, November 20, 2020*, Priyanka Agrawal, Zornitsa Kozareva, Julia Kreutzer, Gerasimos Lampouras, André F. T. Martins, Sujith Ravi, and Andreas Vlachos (Eds.). Association for Computational Linguistics, 43–52. <https://doi.org/10.18653/v1/2020.spnlp-1.6>
- [27] Clément Sage, Alexandre Aussem, Haytham Elghazel, Véronique Eglin, and Jérémy Espinas. 2019. Recurrent Neural Network Approach for Table Field Extraction in Business Documents. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*. IEEE, 1308–1313. <https://doi.org/10.1109/ICDAR.2019.00211>
- [28] Daniel Schuster, Klemens Muthmann, Daniel Esser, Alexander Schill, Michael Berger, Christoph Weidling, Kamil Aliyev, and Andreas Hofmeier. 2013. Intellix - End-User Trained Information Extraction for Document Archiving. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*. IEEE Computer Society, 101–105. <https://doi.org/10.1109/ICDAR.2013.28>
- [29] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1073–1083. <https://doi.org/10.18653/v1/P17-1099>
- [30] Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 11 (2017), 2298–2304. <https://doi.org/10.1109/TPAMI.2016.2646371>
- [31] Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. 2021. MatchVIE: Exploiting Match Relevancy between Entities for Visual Information Extraction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Zhi-Hua Zhou (Ed.). ijcai.org, 1039–1045. <https://doi.org/10.24963/ijcai.2021/144>
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [33] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards Robust Visual Information Extraction in Real World: New Dataset and Novel Solution. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2738–2745. <https://ojs.aaai.org/index.php/AAAI/article/view/16378>
- [34] Jiapeng Wang, Tianwei Wang, Guozhi Tang, Lianwen Jin, Weihong Ma, Kai Ding, and Yichao Huang. 2021. Tag, Copy or Predict: A Unified Weakly-Supervised Learning Framework for Visual Information Extraction using Sequences. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Zhi-Hua Zhou (Ed.). ijcai.org, 1082–1090. <https://doi.org/10.24963/ijcai.2021/150>
- [35] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 1192–1200. <https://doi.org/10.1145/3394486.3403172>
- [36] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florêncio, Cha Zhang, and Furu Wei. 2021. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. *CoRR* abs/2104.08836 (2021). arXiv:2104.08836 <https://arxiv.org/abs/2104.08836>
- [37] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2579–2591. <https://doi.org/10.18653/v1/2021.acl-long.201>
- [38] Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2020. PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 4363–4370. <https://doi.org/10.1109/ICPR48806.2021.9412927>

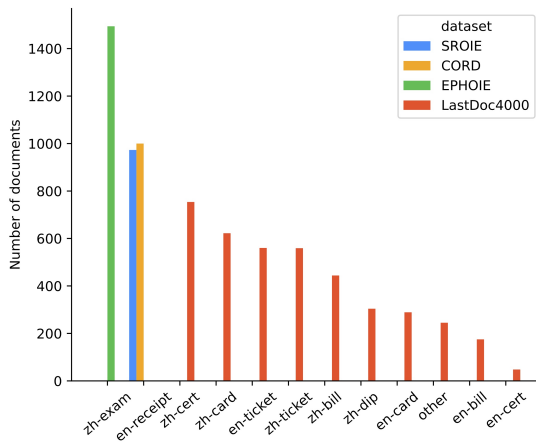


Figure 10: Document quantity of each type. “zh-” is short for Chinese, “en-” for English, “cert” for certificates, “dip” for diploma. It demonstrates that the diversity of LastDoc4000 in document types is superior to other datasets.

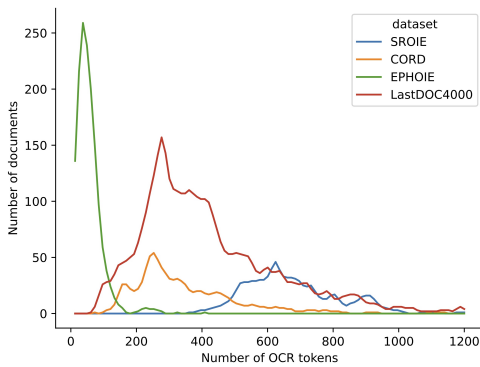


Figure 11: Statistics of word length in different datasets.

Table 8: Ablation study of various augmentation strategies on LastDoc4000 dataset. “Sem”, “Spa” and “Vis” stand for “Semantic”, “Spatial” and “visual” respectively.

Method	Sem	Spa	Vis	Setup-D
				F1(%)
QGN <sub>w/o aug</sub>	✗	✗	✗	79.40
QGN <sub>w/ Sem</sub>	✓	✗	✗	79.88
QGN <sub>w/ Sem&amp;Spa</sub>	✓	✓	✗	80.86
QGN <sub>w/ Sem&amp;Spa&amp;Vis</sub>	✓	✓	✓	<b>81.24</b>

## A LASTDOC4000 DATASET

As stated in the main text, public datasets of DIE have significantly promoted the development of document understanding research. However, most of them focus on limited entities in specific scenarios (e.g., receipts, forms, and examination papers), which lacks the generalization of various scenarios in terms of layouts and keys. In order to address this issue, we introduce a new large-scale

benchmark named LastDoc4000 containing 4,000 images of semi-structured documents. Some examples of LastDoc4000 are shown in Fig. 9.

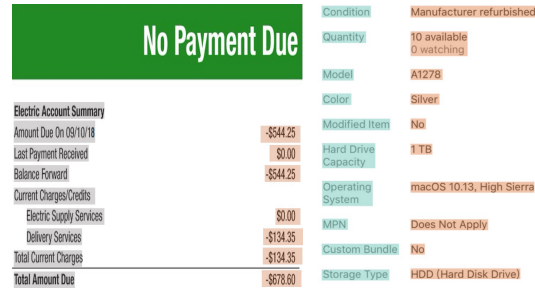


Figure 9: Some examples in LastDoc4000. Key and value entities are highlighted in different colors.

### A.1 Annotation details.

For each document, there are four annotation types for information extraction: 1) Text contents with the bounding box, used as correct inputs of ideal scenes. 2) Recognition results of OCR engine, used as noisy input of the wild scenes. 3) Entity annotation with key-value pairs, used to check the intermediate results of the referenced two-stage extraction method. 4) Structured results of key-value pairs, used as the end-to-end targets, need to be extracted.

### A.2 Statistical analysis.

LastDoc4000 has a large number of documents with various layouts in English and Chinese, aiming to promote DIE in the wild scenes. It contains 4,000 images with 1,511 different layouts, including forms, certificates, bills, and other structured information scenarios. The data is divided into a training set with 2,687 images and a testing set with 1,313 images, respectively. Moreover, to better study the dataset’s diversity, we also collect statistics of the document type and word count. As shown in Fig. 10 and Fig. 11, we can see that LastDoc4000 exhibits more diversity in both document types and content length.

### A.3 Data desensitization.

For the information security issue, we strictly conduct data desensitization processes as follows: 1) Remove the human face and body information in the document. 2) Remove seal information from documents, which mainly appear in enterprise licenses. 3) Remove handwritten notes from documents, which are mainly in the bank receipt scenarios. 4) Replace sensitive field information with synthetic data, including name, gender, company name, telephone number, address, email, etc. 5) Manual review in total.

## B EFFECT OF AUGMENTATION STRATEGY

We validate the effect of proposed data augmentations on QGN, as reported in Tab. 8. With the help of augmentation, a substantial performance boost of margin 1.84% is acquired on the LastDoc4000 dataset. This ablation study proves that each augmentation method is beneficial, and their combination provides a better result.