# Learning Label Dependencies for Visual Information Extraction

**Minghong Yao**, **Liansheng Zhuang***, **Houqiang Li** and **Jiuchang Wei**
University of Science and Technology of China
mhyao1@mail.ustc.edu.cn, {lszhuang, lihq, weijc}@ustc.edu.cn

## Abstract

Visual Information Extraction (VIE), which aims to extract structured information from visually rich document images, has drawn much attention due to its wide applications in document understanding. However, previous methods often treat the VIE task as a sequence labeling problem and ignore the label correlations in the sequence, which may significantly degrade their performance. To address this issue, this paper proposes a novel framework to exploit the potential of label correlations to improve the VIE models' performance. Its key idea is to learn the label dependency of entities, and use it to regularize the label sequence. Specifically, to capture the label dependency of entities, a label transformer is pre-trained to assign a higher likelihood to the label sequence that respects the label patterns of document layouts. During testing stages, an inference transformer is used to predict the label sequence by considering not only the features of each entity but also the likelihood of the label sequence evaluated by the label transformer. Our framework can be combined with existing popular VIE models such as LayoutLM and GeoLayoutLM. Extensive experiments on public datasets have demonstrated the effectiveness of our framework.

## 1 Introduction

The visual information extraction task (VIE) involves extracting texts of multiple key segments from given document images and saving these texts to structured documents. With the acceleration of the digitization process, the VIE task has been regarded as a crucial part of intelligent document processing and is required by many real-world applications in various industries such as finance, medical treatment, and insurance [Cui, 2021]. A VIE task is often divided into two sub-tasks, namely semantic entity recognition (SER, *a.k.a.* entity labeling) and relation extraction (RE, *a.k.a.* entity linking) [Li *et al.*, 2021c; Zhang *et al.*, 2021]. The former aims to extract meaningful text segments (namely entities) from images while the latter is to predict the relations between entities. The past decades have witnessed the progress of VIE. However, it remains an open problem, especially in the wild. In this paper, we focus on the SER sub-task and exploit the patterns in label sequence to improve the performance of existing VIE models.

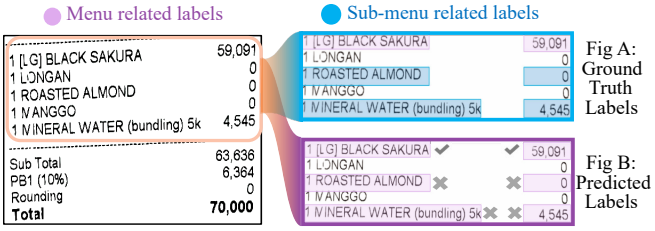


Figure 1: A sample image selected from the CORD dataset. In this image, two different kinds of labels, "menu" and "sub-menu", are not distinguishable based on text, image, and layout features. A common mistake made by existing models is that they confuse "sub-menu" related labels with "menu" related labels. By examining more images in the CORD dataset, we find that "menu" related labels are always followed by "sub-menu" related labels. This pattern is observable only in the label sequence. Better viewed in color.

Generally, the sub-task of semantic entity recognition (SER) is formalized as a sequence tagging problem and how to capture effectively the dependency between entities has been the focus. Document layouts, which reflect the geometric relationships between entities, have shown consistent benefits in helping capture the dependency. So, some popular methods directly learn the layout embedding (such as LayoutLM [Xu *et al.*, 2020]), and have proven to be effective for SER even without using other features (such as text) in [Wang *et al.*, 2022]. Another popular methods build various graphs to capture the entities dependency, where each node represents an entity and each edge represents the distance between two entities. Graph neural networks (GNNs) are used to fuse layout features and text features [Yu *et al.*, 2020; Liu *et al.*, 2019; Tang *et al.*, 2021] between different entities along the graphs. By considering the feature-level dependency of entities, existing methods have achieved an impressive performance in public data sets.

Besides the feature-level dependency of entities, the label-level dependency of entities is also important for VIE tasks. Label dependencies are defined as correlations between la-

*Corresponding Author

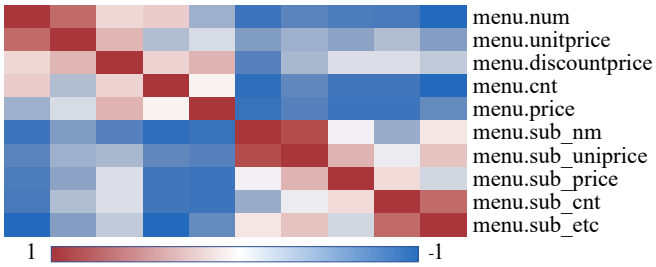


Figure 2: Correlation matrix between "menu" and "sub-menu" related labels in the CORD dataset. A higher value in each cell indicates a higher correlation. We find the labels of this dataset are spatially correlated by the following sampling procedures. First, we generate 100 sampling bounding boxes randomly in the 2D plane for each document image. Then, we collect the labels of text segments whose bounding boxes overlap with the same generated sampling box and increase the co-occurrence count between all pairs of the collected labels. Finally, we apply the chi-squared test to compute the correlation score using the co-occurrence matrix obtained in the previous two steps. Better viewed in color.

bels, which means that two labels are not statistically independent. Given a document image, the entity labels are distributed in a plane with a explicit pattern. Arising from these patterns, the labels of entities have a high correlation, which can be used to improve performance. Fig. 1 shows a receipt of a restaurant in CORD dataset [Park *et al.*, 2019]. Existing VIE models often fail to disguish the entity categories of pink text segments ("menu") from the blue text segments ("sub-menu") based on multimodal features. For example, LayoutLMv3 [Huang *et al.*, 2022] incorrectly predicted "sub-menu" as "menu" as shown in subfigure B of Fig. 1. In fact, a "menu" entity is always followed by several "sub-menu" entities in a receipt of restaurants. If we can use the label dependency, it's reasonable to expect LayoutLMv3 to correctly distinguish these entities labels.

What's more, the label dependency occurs in large numbers in VIE tasks. Indeed, statistical testing shows that correlated labels tend to occupy the same space in the 2D plane. Fig. 2 shows the testing result for the CORD dataset. More results on the other datasets can be found in the appendixes. This phenomenon exists a lot in document images because entities need to be arranged in the 2D plane according to certain patterns (such as "key-value" pairs) for better readability. Therefore, the label dependency of entities provides a complementary yet important cue for VIE tasks. However, most existing methods ignore the label dependency, which significantly degrades their performance.

Meanwhile, it is non-trivial to learn the label dependency because the short-range dependency between entities in a 2D plane can become a long-range dependency in the process of text serialization [Wang *et al.*, 2021; Gu *et al.*, 2022]. This makes conventional methods such as linear-chain Conditional Random Fields (CRFs) used in many works unsuitable in the VIE task [Huang *et al.*, 2015; Rrubaa and Amaresan, 2018; Yu *et al.*, 2020]. Linear-chain CRFs make two key assumptions: first, that the current label depends only on the previous label; second, that the transition probabilities between different labels are independent of observed features. However, these assumptions are not true in the document image domain. Moreover, the computational complexity of higher-order CRFs grows exponentially. Therefore, it is still an open problem to learn long-range label dependency in VIE tasks.

Inspired by above insights, this paper proposes a fine-tuning framework to boost the performance of existing VIE models. The core idea is to add a penalty loss during fine-tuning such that the predicted label sequence is regularized by the label dependency. Specifically, the negative loglikelihood of a label sequence is used as the penalty loss. To this end, we introduce a label transformer and pre-train it by the next label prediction task using the ground truth label sequences. During fine-tuning, this label transformer is expected to assign a higher likelihood to the label sequence that respects the label patterns of document layouts. Since the landscape of our penalty term is highly non-convex, it's non-trivial to find the correct label sequence that can minimize this term. To address this technical difficulty, we turn to the help of adversarial training, i.e., a second model (namely the inference transformer) is trained to predict the optimal label sequence that minimizes the penalty loss evaluated by the label transformer. The reference transformer feeds on the sequence of features and labels. We note that the initial label sequence is random and this randomness can cause the inference transformer to predict unstable label sequences. To address this issue, we propose to fit the label sequence by the fixed point of inference transformer. Specifically, the inference transformer model starts from a random label sequence and refines the label sequence in parallel according to the feature sequence. We iterate this process until the label sequence converges to a fixed point.

Our contributions are summarized as follows:

- This paper proposes a fine-tuning framework to boost the performance of pre-trained multimodal models. In the framework, a label transformer is introduced to learn long-range label dependency, and an inference transformer is introduced to predict the label sequence regularized by the label dependency.
- To address initial randomness in label sequences, the inference transformer iteratively refines the label sequence until it reaches a stable fixed point.
- Extensive experiments on public dataset show that label dependency can boost the performance of existing popular VIE models.

## 2 Related Works

The VIE task has attracted the attention of many researchers in recent years. Most previous works model the sets of text segments either as a graph or as a sequence. Many works [Qian *et al.*, 2019; Liu *et al.*, 2019; Yu *et al.*, 2020; Tang *et al.*, 2021; Cheng *et al.*, 2020; Yao *et al.*, 2021] construct a document graph, using text features as node features and the relative spatial features of segments as edge features. The typical works of this line include PICK [Yu *et al.*, 2020] and MatchVIE [Tang *et al.*, 2021]. These approaches employed Graph Neural Networks (GNNs) to generate text embeddings from layout features, aiming to understand the key-value relationships. By passing messages between nodes,

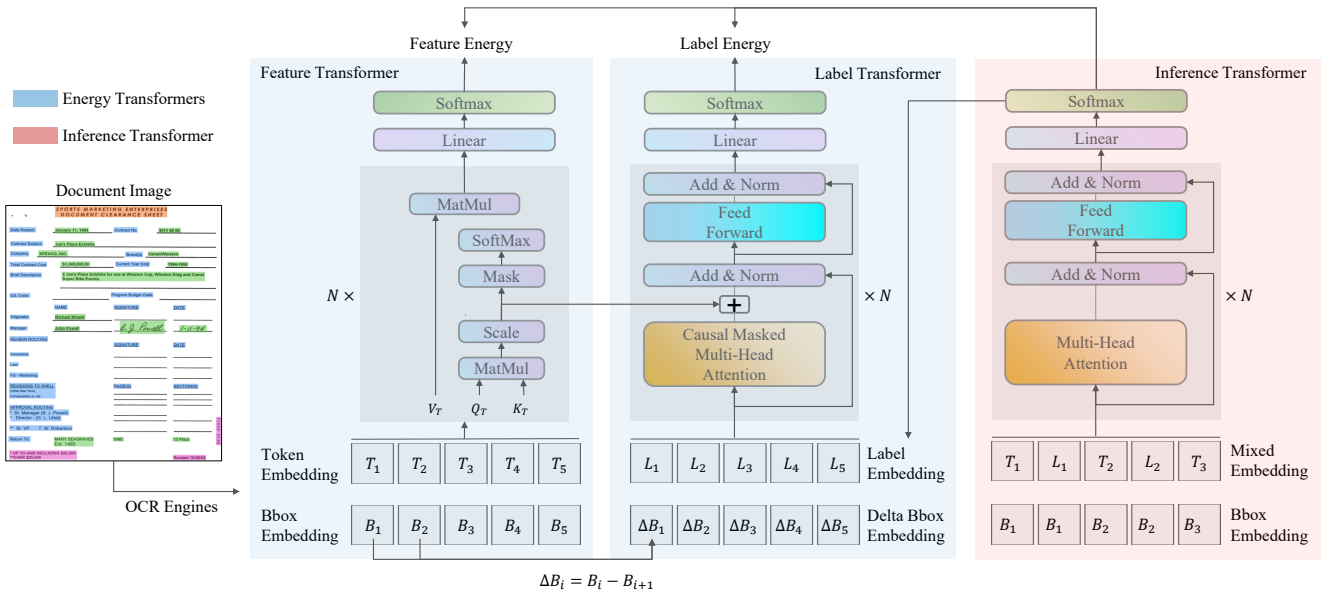


Figure 3: Our framework consists of three neural networks, including feature transformer, label transformer, and inference transformer. Only the inference transformer is used to do inference during the test. Feature and label transformers are used to calculate feature and label energy, which will help train the inference transformer during training. They are termed energy transformers as indicated by the legends.

these models were capable of comprehending the entire layout and the specific distribution of each text segment. The advantage of graph-based models is that they have low computational complexity. However, the state-of-the-art performances are achieved by the sequence-based models.

The core of sequence-based models relies on how to leverage the power of pre-training to align multimodal features and then learn discriminative features during fine-tuning. The early baseline model applies a classical bi-directional long short-term memory network (Bi-LSTM) [Huang *et al.*, 2015] to the sequence of text segments and marks where the interested categories of entities are by predicting BIO labels for each segment [Yu *et al.*, 2020]. The problem is that two text segments that are close to each other in the 2D plane can be separated by a large distance after serialization [Wang *et al.*, 2021; Gu *et al.*, 2022]. This causes difficulty for sequential modeling because of the abrupt shift in thematic focus. To this end, more advanced sequential models (mostly Transformers [Vaswani *et al.*, 2017]) also feed on the coordinates of each text segment or even the original images to mitigate this problem. Inspired by the breakthrough of pretraining technology [Devlin *et al.*, 2019], many works have been focusing on the alignment of multimodal features (textual, layout, and images) in pretrainin [Xu *et al.*, 2020; Xu *et al.*, 2021; Li *et al.*, 2021a; Appalaraju *et al.*, 2021; Li *et al.*, 2021b; Huang *et al.*, 2022; Luo *et al.*, 2023]. Then the models learn discriminative features for each text segment in fine-tuning such that the predicted BIO label sequence is accurate. However, these methods cannot work well when facing ambiguous multimodal features during fine-tuning.

Therefore, it is important to model the label dependencies explicitly for VIE. The models [Huang *et al.*, 2015; Yu *et al.*, 2020; Zhang *et al.*, 2021] in VIE mainly apply the linear-chain CRFs developed in structured learning [Nowozin *et al.*, 2011]. However, CRFs cannot work well when facing long-range label dependencies. To this end, the SPEN model, which uses neural networks to learn energy functions, is proposed in [Belanger and McCallum, 2016]. Sooner, [Tu and Gimpel, 2018] proposes two adversarial neural networks where one evaluates the plausibleness of a label sequence and the other one learns to fool the evaluator.

## 3 Our Model

### 3.1 Problem Setup

Let $\mathbf{x} = \{x_n\}_{n=1,\cdots,N}$ be a sequence of features. They are the multimodal features of serialized text, including textual, layout, and image features. $N$ is the sequence length. We aim to predict the label for each $x_n$. The ground truth label sequence is denoted as $\mathbf{y} = \{y_n\}_{n=1,\cdots,N}$. The predicted label sequence is denoted as $\hat{\mathbf{y}} = \{\hat{y}_n\}_{n=1,\cdots,N}$. Let $M$ be the size of the label set. Mathematically speaking, we aim to build a model that learns the conditional probability distribution of a label sequence $\mathbf{y}$ given a feature sequence $\mathbf{x}$ from the training dataset. Let $P(\mathbf{y}|\mathbf{x})$ be this distribution. By considering the long-range dependency between labels, $P(\mathbf{y}|\mathbf{x})$ is precisely factorized as:

$$P(\mathbf{y}|\mathbf{x}) = P(y_1|\mathbf{x})P(y_2|y_1;\mathbf{x})\cdots P(y_n|y_{1,\cdots,n-1};\mathbf{x}).$$

### 3.2 The Framework of Our Model

Inspired by the assumption made in the structured learning area, we choose the exponential family distribution to parameterize the conditional probabilities of each label, $P(y_i|y_{1,\cdots,i-1};\mathbf{x})$. In the exponential family distribution, a parameterized function assigns a real value to any collection of random variables. This real value serves as the unnormalized probability of these random variables. This function is

called an energy function. To gain the normalization constant, one takes the exponential value of an energy function and then summarizes it over the sample space.

$P(\mathbf{y}|\mathbf{x})$ is modeled by energy functions. According to the rule of Bayes, $P(\mathbf{y}|\mathbf{x})$ is factorized as:

$$P(\mathbf{y}|\mathbf{x}) = P_{\mathbf{x}}(y_1)\frac{P_{\mathbf{x}}(y_{1,2})P_{\mathbf{x}}(y_2)\cdots P_{\mathbf{x}}(y_{1,\cdots,n})P_{\mathbf{x}}(y_n)}{P_{\mathbf{x}}(y_1)P_{\mathbf{x}}(y_2)\cdots P_{\mathbf{x}}(y_{1,\cdots,n-1})P_{\mathbf{x}}(y_n)}.$$

Take the logarithm value on both sides and rearrange terms:

$$\log P(\mathbf{y}|\mathbf{x}) = \sum_{n=1}^{N}\log P_{\mathbf{x}}(y_n) + \log\frac{P_{\mathbf{x}}(y_{1,\cdots,n})}{P_{\mathbf{x}}(y_{1,\cdots,n-1})P_{\mathbf{x}}(y_n)}.$$

The first summation is called feature energy and it measures the log-likelihood of $\mathbf{y}$ without considering the label dependencies. The second term is called label energy and it specifically models the influence of label dependencies. The exponential function of $P(\mathbf{y}|\mathbf{x})$ will cancel the logarithm function. This means the log-likelihood is reduced to the summation of feature and label energy. We illustrate how to calculate feature and label energy in Fig.3, which also shows the framework of our model.

Let $F_\phi(\cdot)$ denote the feature transformer and $L_\phi(\cdot)$ denote the label transformer, where $\phi$ is the parameter of them. Then, for the $n$th input feature, its corresponding logits vector predicted by $F_\phi$ is denoted as $F_\phi(\mathbf{x})_n \in \mathcal{R}^M$, where the $i$th element of $F_\phi(\mathbf{x})_n$ is the possibility of $x_n$ having $i$th label. Similarly, let $L_\phi(\mathbf{y})_{n-1} \in \mathcal{R}^M$ denotes logits vector predicted by $L_\phi(\cdot)$ for $n$th input feature given the previous labels $\{y_0, \cdots, y_{n-1}\}$. Let $I_\varphi(\cdot)$ denote the inference transformer. It feeds on both feature and predicted label sequences and predicts a new label sequence, i.e., $I_\varphi(\mathbf{x}, \hat{\mathbf{y}})_n \in \mathcal{R}^M$ represents the logits vector of $n$th input feature and $\hat{\mathbf{y}}$ is the predicted label sequence. We can now introduce the following feature energy $E_{feat}(F_\phi(\mathbf{x}), I_\varphi(\mathbf{x}, \hat{\mathbf{y}}))$, label energy $E_{label}(L_\phi(\mathbf{x}), I_\varphi(\mathbf{x}, \hat{\mathbf{y}}))$ and energy function $E_{total}$:

$$\begin{aligned} E_{feat}(\mathbf{x}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}})) &= \textstyle\sum_{n=1}^{N} F_\phi(\mathbf{x})_n^\top I_\varphi(\mathbf{x}, \hat{\mathbf{y}})_n, \\ E_{label}(\mathbf{x}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}})) &= \textstyle\sum_{n=1}^{N} L_\phi(\hat{\mathbf{y}})_n^\top I_\varphi(\mathbf{x}, \hat{\mathbf{y}})_n, \\ E_{total}(\mathbf{x}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}})) &= -(E_{feat} + E_{label}). \end{aligned} \tag{1}$$

### 3.3 Feature Transformer

We select open-source pre-trained models as feature transformers, including LiLT [Wang *et al.*, 2022], LayoutLM series [Xu *et al.*, 2020; Xu *et al.*, 2021; Huang *et al.*, 2022], and GeoLayoutLM [Luo *et al.*, 2023]. In Fig. 3, the input for the feature transformer illustrates only text and layout inputs, omitting the input for images. More general types of transformers are also permitted if they accept text, layout, and images as input feature sequences and calculate attention scores between any two tokens based on these features. It is worth emphasizing that the commonly used type of transformer here is the encoder-only type, where any two tokens can mutually attend to each other without the restriction of causal masking. We include the final linear layer and softmax layer in the feature transformer. the softmax layer transforms the output energy into bounded numerical values, facilitating the inference transformer to converge to the correct label sequence.

### 3.4 Label Transformer

We choose a decoder-only type of transformer to learn long-range dependencies between labels. This involves applying a causal mask when calculating attention scores between labels to ensure that each token can only see itself and the preceding labels when predicting the next label. As a result, the likelihood of the $n$th token having a certain label depends on the logits vectors output by the $n-1$th token and before. Due to the presence of a causal mask, the model is compelled to learn how labels transition during training and avoid simply memorizing the input label sequence.

In the domain of document images, when there is a sudden change in layout, such as a line break or the insertion of a tab character, the upcoming text will likely have a different functional role. Similarly, abrupt changes in width and height imply changes in font size, which may also result in changes to text labels. To detect the changes, we first embed the layout of each token and then subtract the layout embedding of the current position from that of the next position.

Furthermore, we observe that it is beneficial to add attention scores from the feature transformer into the label transformer's attention scores. This is reasonable because the attention scores in the feature transformer allow the label transformer to focus on a limited set of labels in the preceding context when predicting the next label.

Finally, we will fine-tune the feature transformer initially following the conventional fine-tuning approach. Subsequently, we will pre-train the label transformer using the real label sequences from the training set. During this process, we leverage the attention scores produced by the feature transformer (with the Feature Transformer being frozen).

### 3.5 Inference Transformer

As observed in [Dupont *et al.*, 2018], embedding the label sequence into the feature sequence has been shown to effectively enhance model performance. To address the challenge of not knowing the ground truth label sequence initially, we introduce the concept of fixed-point fitting. Inspired by the work described in [Bai *et al.*, 2019; Bai *et al.*, 2020], the inference transformer predicts the label sequence by iterative denoising from an initial random label sequence. The output of the inference transformer is used as its input in the next iteration, formally expressed as:

$$\hat{\mathbf{y}}^{t+1} = I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^t). \tag{2}$$

In the above equation, the superscript $t$ denotes the result of the $t$-th iteration. As shown in Fig 3, we insert the label embedding of $\hat{y}_i$ after each feature embedding $x_i$ and call them mixed embedding. This iteration stops when a fixed point, $\hat{\mathbf{y}}^*$, is reached. The fixed point $\hat{\mathbf{y}}^*$ satisfies:

$$\hat{\mathbf{y}}^* = I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^*). \tag{3}$$

We consider this iterative process as stepwise denoising of the label sequence, allowing the model to repeatedly assess whether the current label sequence can be improved until further improvement is no longer possible. To ensure the existence of a fixed point, we use a normalized logit vector for each $\hat{y}_i$. We can exploit any black-box root-finding algorithm

**Algorithm 1** Training procedures.

**Require:** Data $\{\mathbf{x}, \mathbf{y}\}$, Transformers $F_\phi$, $L_\phi$, and $I_\varphi$, Inner/Outer Steps $T_{inner}, T_{outer}$

**for** $t \in T_{outer}$ iterations **do**
  Obtain sample $\mathbf{x}, \mathbf{y}$ from dataset
  Freeze $F_\phi$, $L_\phi$ and fine-tune $I_\varphi$
  **for** $p \in T_{inner}$ iterations **do**
    Find $\hat{\mathbf{y}}^* = I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^*)$
    Compute $\mathcal{L}_{Aux}\left[\mathbf{y}, E_{total}(\mathbf{x}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^*))\right]$
    $\varphi \leftarrow \varphi - gradient(\mathcal{L}_{Aux}, \varphi)$
  Freeze $I_\varphi$ and fine-tune $F_\phi$, $L_\phi$
  Compute $\mathcal{L}_{Prim}\left[\mathbf{y}, E_{total}(\mathbf{x}, \mathbf{y})\right]$
  $\phi \leftarrow \phi - gradient(\mathcal{L}_{Prim}, \phi)$

to find the fixed point by solving the root of the following equation, where $RootFind$ is the root-finding algorithm:

$$\hat{\mathbf{y}}^* = RootFind(I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^t) - \hat{\mathbf{y}}^{t+1}). \tag{4}$$

### 3.6 Training and Inference

In our framework, the energy score produced by the feature and label transformers can be interpreted as the emission energy and label transition energy in linear-chain CRFs. Traversing all possible label sequences is an essential step in the learning and inference process of such algorithms. Linear-chain CRFs can use two linear complexity algorithms, the forward-backward algorithm, and the Viterbi algorithm, to estimate the partition function and the optimal label sequence, respectively. The core of designing training and inference algorithms for our framwork lies in alleviating the difficulties of traversing all possible label sequences.

For a given input feature sequence, the label sequence that minimizes the energy function $E_{total}$ is considered the optimal prediction result. To alleviate the difficulties of iterating over all possible label sequences, the inference transformer is trained to achieve the minimum value of $E_{total}$. In the meanwhile, the energy transformers should learn to assign the lowest energy to the ground truth label sequence during training. Put it all together, the following bi-level optimization problem is introduced:

$$\begin{aligned} \varphi = &\quad \arg\min_{\varphi} \mathcal{L}_{Aux}\left[\mathbf{y}, E_{total}(\mathbf{x}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^*))\right], \\ \phi = &\quad \arg\min_{\phi} \mathcal{L}_{Prim}\left[\mathbf{y}, E_{total}(\mathbf{x}, \mathbf{y})\right]. \end{aligned} \tag{5}$$

where $\mathcal{L}_{Prim}$ will be designed to let the ground truth label sequence to achieve the minimal energy and $\mathcal{L}_{Aux}$ will be designed to help the inference transformer learn to minimize the energy function. The training procedures are described in the algorithm 1.

**Training loss.** In the initial phase of training, $E_{total}$ has not yet learned to score the true label sequence correctly. Thus the inference transformer will receive poor supervision. To alleviate this issue, we add direct supervision to the inference transformer using $\mathbf{y}$. We use the cross entropy (CE) loss to add this supervision:

$$\mathcal{L}_{Aux} = CE(\mathbf{y}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^*)) + \lambda E_{total}(\mathbf{x}, I_\varphi(\mathbf{x}, \hat{\mathbf{y}}^*))). \tag{6}$$

| Dataset | # Images | #Types |
|---|---|---|
| FUNSD | Train 149, Test 50 | Forms |
| CORD | Train 800, Test 100 | Receipts |
| SROIE | Train 626, Test 276 | Receipts |

Table 1: Labels, types and the number of images for each dataset.

where $\lambda$ is a hyperparameter. We use a contrastive learning loss to train the energy transformers. For each ground truth label sequence $\mathbf{y}$, $K$ random label sequences $\hat{\mathbf{y}}$ are sampled. The loss reads:

$$\mathcal{L}_{Prim} = -\log \frac{\exp(-E_{total,\phi}(\mathbf{x}, \mathbf{y}))}{\sum_{k=0}^{K} \exp(-E_{total,\phi}(\mathbf{x}, \hat{\mathbf{y}}))}. \tag{7}$$

## 4 Experiments

### 4.1 Datasets

Table 1 lists three VIE benchmark datasets: FUNSD [Jaume *et al.*, 2019], CORD [Park *et al.*, 2019], and SROIE [Huang *et al.*, 2019]. The FUNSD dataset is designed for form understanding. The dataset encompasses various document types, including market reports, advertisements, academic reports, etc. The CORD dataset collects 1,000 Indonesian receipts from shops and restaurants and annotates them with 5 superclass and 42 subclass labels. In the SROIE dataset, each receipt image contains four key text fields, such as total cost, goods name, unit price, etc.

### 4.2 Baseline

In the VIE task, the current state-of-the-art (SOTA) models are mostly based on the transformer architecture. We selected LiLT [Wang *et al.*, 2022], LayoutLM series [Xu *et al.*, 2020; Xu *et al.*, 2021; Huang *et al.*, 2022], and GeoLayoutLM [Luo *et al.*, 2023] as feature transformers. As baseline models, we also presented their F1 scores in the paper before considering structural learning. Furthermore, to assess whether our approach learned long-range label dependencies, we applied the linear-chain CRF model to these baseline models for comparison. After hyperparameter searching using Bayesian optimization, we trained the label transition matrix in the linear-chain CRF with a learning rate of 1e-3.

### 4.3 Implementation Details

**Details of Fine-tuning Energy Transformers.** The selected feature transformers have reported in papers the fine-tuning hyperparameters that achieve optimal performance. Therefore, we fine-tune them using the reported hyperparameters. After freezing their parameters, we utilize their attention scores to assist the Label Transformer in pre-training a language model for labeling on public datasets. The AdamW optimizer is employed for fine-tuning, with an initial learning rate of 5e-5 and a linear decay learning rate scheduler.

**Details of Fine-tuning Inference Transformers.** We employed the AdamW optimizer to train it, with a learning rate set to 5e-5 and a batch size of 1. For the fixed-point search, we utilized the Broyden [Broyden, 1965] method, with a maximum iteration step limit set to 10.

| Models | Original F1 | With CRF | With Ours |
|---|---|---|---|
| LiLT-base | 88.4 | 89.0 (↑) | 90.1 (↑) |
| LiLT-large | 90 | 90.5 (↑) | 91.2 (↑) |
| LayoutLM-base | 78.7 | 79.3 (↑) | 80.5 (↑) |
| LayoutLM-large | 79 | 79.5 (↑) | 83.2 (↑) |
| LayoutLMv2-base | 82.8 | 83.4 (↑) | 86.7 (↑) |
| LayoutLMv2-large | 84.2 | 84.7 (↑) | 88.5 (↑) |
| LayoutLMv3-base | 90.3 | 91 (↑) | 91.7 (↑) |
| LayoutLMv3-large | 92.1 | 92.2 (↑) | 92.7 (↑) |
| GeoLayoutLM | 92.3 | 92.1 (↓) | 93.1 (↑) |

Table 2: Results of FUNSD Dataset

| Models | Original F1 | With CRF | With Ours |
|---|---|---|---|
| LiLT-base | 95.11 | 95.91 (↑) | 96.3 (↑) |
| LiLT-large | 96.07 | 96.48 (↑) | 96.86 (↑) |
| LayoutLM-base | 94.6 | 94.9 (↑) | 965.4 (↑) |
| LayoutLM-large | 95.7 | 96.24 (↑) | 97.1 (↑) |
| LayoutLMv2-base | 96.25 | 96.8 (↑) | 97.67 (↑) |
| LayoutLMv2-large | 96.61 | 97.12 (↑) | 97.99 (↑) |

Table 3: Results of SROIE Dataset

## 4.4 Comparison with the SOTAs

Given the results presented across Tables 2, 4, and 3, which show the performance of various models on the FUNSD, CORD, and SROIE datasets, we can see our proposed method has achieved the state-of-the-art results.

The proposed methods have led to a consistent improvement in F1 scores across different datasets, indicating a robust adaptability and effectiveness of the approach. Specifically, the application of our methods has resulted in statistically significant performance boosts as denoted by the upward arrows in the tables, which suggests that the improvements are not only consistent but also substantial.

For the FUNSD dataset, which is known for its challenging forms and diverse layouts, our methods have resulted in marked performance improvements. The GeoLayoutLM model, in particular, has shown exceptional gains, achieving a 93.1 F1 score which surpasses its original F1 score by a significant margin. This enhancement is indicative of the method's efficacy.

Turning to the CORD dataset, which focuses on receipt understanding, the results are even more pronounced. Here, GeoLayoutLM again stands out, with a notable increase in F1 score to 98.2 when our methods are applied. The CORD dataset is particularly demanding due to the irregular formats and dense information presented in receipts. The improvements here underscore our method's capability to discern and interpret detailed information within noisy and unstructured data environments.

The SROIE dataset, comprising receipt data similar to CORD but with its own unique challenges, again shows a similar trend of improvement. The introduction of our methods has pushed the boundaries of model performance, with the F1 score for LayoutLMv2-large reaching an impressive

| Models | Original F1 | With CRF | With Ours |
|---|---|---|---|
| LiLT-base | 95.11 | 96.10 (↑) | 97.1 (↑) |
| LiLT-large | 96.07 | 96.18 (↑) | 97.36 (↑) |
| LayoutLM-base | 94.5 | 94.9 (↑) | 95.4 (↑) |
| LayoutLM-large | 95.18 | 95.24 (↑) | 96.1 (↑) |
| LayoutLMv2-base | 95.37 | 95.43 (↑) | 96.67 (↑) |
| LayoutLMv2-large | 96.01 | 96.10 (↑) | 96.99 (↑) |
| LayoutLMv3-base | 96.6 | 96.66 (↑) | 97.38 (↑) |
| LayoutLMv3-large | 97.46 | 97.51 (↑) | 97.83 (↑) |
| GeoLayoutLM | 97.97 | 97.94 (↓) | 98.2 (↑) |

Table 4: Results of CORD Dataset

| # | $L_\phi(\cdot)$ | $\Delta B$ | Mix | FIP | F1 | Precision | Recall |
|---|---|---|---|---|---|---|---|
| 1a | ✓ | × | × | × | 88.7 | 88.2 | 89.3 |
| 1b | ✓ | ✓ | × | × | 89.3 | 88.7 | 90 |
| 1c | × | × | ✓ | × | 88.4 | 88.9 | 87.9 |
| 1d | × | × | ✓ | ✓ | 88.6 | 88.9 | 88.3 |
| 2a | ✓ | × | ✓ | ✓ | 89.5 | 90.5 | 88.6 |
| 2b | ✓ | ✓ | ✓ | × | 90.0 | 90.0 | 90.0 |
| 3 | ✓ | ✓ | ✓ | ✓ | 90.1 | 91.0 | 89.3 |

Table 5: Testing the impact of different components on the Funsd Dataset with LiLT-base Feature Transformer. In the table head, "$L_\phi(\cdot)$" represents the label transformer; "$\Delta B$" represents the delta box embedding; "Mix" represents mixed embedding; "Fix" represents fixed point iteration.

97.99. This further demonstrates the method's strength in extracting text and understanding structure, which is crucial for tasks such as information extraction from receipts where precision is paramount.

Across all datasets, it is evident that the incorporation of CRF has provided a boost to model performance. However, the additional enhancements from our methods lead to even higher F1 scores. This is particularly true for the larger models, which already have a high baseline performance due to their increased complexity and capacity to capture fine-grained patterns in data. The fact that our methods can still contribute to significant gains in performance speaks to the innovative nature of the approaches we have implemented.

## 4.5 Abliation Study

To better understand the effectiveness of different components in our framework, we conduct ablation studies for the label transformer, delta box embedding, mixed embedding, and fixed point iteration. We select the Funsd dataset and LiLT-base Feature Transformer to conduct experiments.

The results from Table 5 present a clear indication of the effectiveness of the proposed components in VIE tasks. The ablation study shows how each element, including the label transformer, delta box embedding, mixed embedding, and fixed point iteration, contributes to the overall performance.

Notably, the label transformer emerges as a cornerstone component, with its presence being essential for achieving high F1, precision, and recall scores. Configurations without the label transformer (1c and 1d) lag behind those that include it (1a, 1b, 2a, 2b, and 3). This underscores its critical role in our framework.

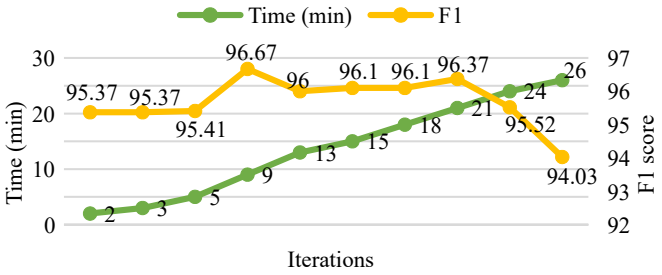


Figure 4: Performance and computation complexity. Better viewed in color.

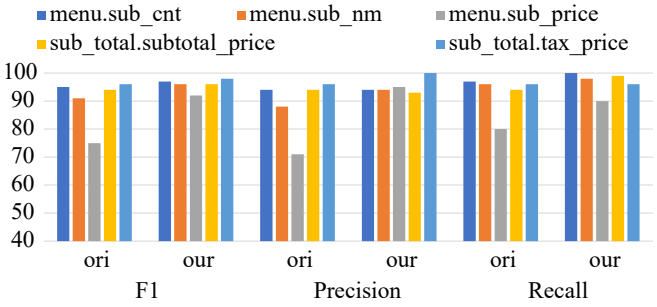


Figure 5: Comparison of performance on 5 selected labels in the CORD dataset. "ori" means the original model. Better viewed in color.

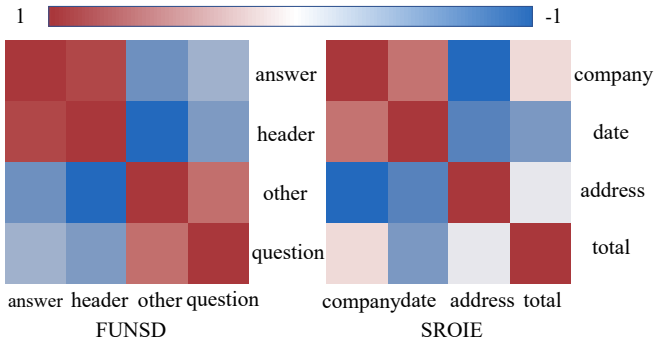


Figure 6: Label dependencies in the FUNSD and SROIE dataset. Better viewed in color.

The delta box embedding also proves to be a significant factor. Its inclusion in configurations 1b and 3 results in a noticeable improvement in the performance, indicating its effectiveness in capturing spatial relationships within the data, which is pivotal for the label transformer.

In the configuration 1c and 2b, we stop the fixed point iteration in the second iteration. As a result, the mixed embedding component, although not as influential as the delta box embedding when used alone, shows its strength when combined with other components. The combination of mix embedding and delta box embedding in configuration 2b yields an improved F1 score over configuration 1b, suggesting that the mix embedding contributes valuable contextual information that complements the spatial insights provided by the delta box embedding. The F1 score of 1d is higher than 1c by 0.2 and configuration 3 is higher than 2b by 0.1. This consistent performance gain proves that fixed point iteration is another vital contributor.

The culmination of all components in configuration 3 results in the highest F1 score (90.1), precision (91), and a strong recall (89.3). This configuration underscores the synergistic effect of combining these components, leading to a model that is precise in its predictions. In summary, the ablation study validates the proposed components, each contributing uniquely to the VIE task. The incremental improvements seen with the addition of each component confirm their value.

### 4.6 Hyperparameter Searching

We notice that the number of inner iterations $T_{inner}$ described in Algorithm 1 has a large impact on both the computational complexity and performance. To achieve balance, we apply grid searching to find the best $T_{inner}$. We conduct these experiments on the CORD dataset using the LayoutLMv2-base.

As shown in Fig. 4, the performance of our framework is not proportional to the iteration numbers. Large iterations lead to worse performance and high computation complexity. Therefore, we set $T_{inner}$ to be 4 during training and testing.

### 4.7 Case Study

In the case study, we addressed a specific issue with the LayoutLMv3 model, which was struggling to accurately classify entities in a document where text, layout, and image features were similar but required distinct labels. As shown in Fig. 5, we select five hard labels in the CORD dataset to varify our framework. The result was a significant improvement in the model's performance. In these test cases, where LayoutLMv3 initially had a high error rate, our enhanced model correctly identified and differentiated the entities, effectively rectifying the previous inaccuracies.

## 5 Conclusions

In conclusion, this paper has demonstrated the efficacy of a novel document understanding framework through rigorous evaluation on the FUNSD, CORD, and SROIE datasets. The introduction of key components such as the label transformer, mix embedding, fixed-point iteration, and delta box embedding, has significantly enhanced model performance. The consistent performance improvements across diverse datasets attest to the robustness and adaptability of the approach. With its demonstrated superiority, the proposed methodology establishes a new state-of-the-art in document understanding, promising to influence a wide array of applications in the realm of automated document processing.

## Appendix

We follow the procedures described in Fig. 2. Fig. 6 shows the results. In the FUNSD dataset, we generate sampling boxes that are 15 times bigger than the average bounding box. In the CORD dataset, sampling boxes are 20 times bigger than the average value. The text segments are sparse in FUNSD and SROIE datasets.

## Acknowledgements

This work was supported in part to Dr. Liansheng Zhuang by NSFC under contract No. U20B2070 and No. 61976199.

## References

[Appalaraju *et al.*, 2021] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[Bai *et al.*, 2019] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.

[Bai *et al.*, 2020] Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. *Advances in Neural Information Processing Systems*, 2020.

[Belanger and McCallum, 2016] David Belanger and Andrew McCallum. Structured prediction energy networks. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[Broyden, 1965] C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 1965.

[Cheng *et al.*, 2020] Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. One-shot text field labeling using attention and belief propagation for structure information extraction. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[Cui, 2021] Lei Cui. Document ai: Benchmarks, models and applications (presentation@ icdar 2021). In *DIL workshop in ICDAR*, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, 2019.

[Dupont *et al.*, 2018] Yoann Dupont, Marco Dinarelli, and Isabelle Tellier. Label-dependencies aware recurrent neural networks. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I 18*, 2018.

[Gu *et al.*, 2022] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[Huang *et al.*, 2019] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition*, 2019.

[Huang *et al.*, 2022] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[Jaume *et al.*, 2019] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019.

[Li *et al.*, 2021a] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 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 (Volume 1: Long Papers)*, 2021.

[Li *et al.*, 2021b] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[Li *et al.*, 2021c] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. Structured text understanding with multi-modal transformers. In *MM '21: ACM Multimedia Conference*, 2021.

[Liu *et al.*, 2019] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, 2019.

[Luo *et al.*, 2023] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Nowozin *et al.*, 2011] Sebastian Nowozin, Christoph H Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 2011.

[Park *et al.*, 2019] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. CORD: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[Qian *et al.*, 2019] Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 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, Volume 1 (Long and Short Papers)*, 2019.

[Rrubaa and Amaresan, 2018] Panchendrarajan Rrubaa and Aravindh Amaresan. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, 2018.

[Tang *et al.*, 2021] Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. Matchvie: Exploiting match relevancy between entities for visual information extraction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[Tu and Gimpel, 2018] Lifu Tu and Kevin Gimpel. Learning approximate inference networks for structured prediction. In *International Conference on Learning Representations*, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 2017.

[Wang *et al.*, 2021] Yiheng Wang, Zilong nd Xu, Lei Cui, Jingbo Shang, and Furu Wei. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[Wang *et al.*, 2022] Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*, 2022.

[Xu *et al.*, 2020] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[Xu *et al.*, 2021] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 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 (Volume 1: Long Papers)*, 2021.

[Yao *et al.*, 2021] Minghong Yao, Zhiguang Liu, Liangwei Wang, Houqiang Li, and Liansheng Zhuang. One-shot key information extraction from document with deep partial graph matching. *arXiv preprint arXiv:2109.13967*, 2021.

[Yu *et al.*, 2020] Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2020.

[Zhang *et al.*, 2021] Yue Zhang, Zhang Bo, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. Entity relation extraction as dependency parsing in visually rich documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.