Enforcing Graph Structures to Enhance Key Information Extraction in Document Analysis

Rajashree Majumder[1], Zhewei Wang[1], Ye Yue[1], Mukut Kalita[2] and Jundong Liu[1,*]
[1]*School of Electrical Engineering and Computer Science, Ohio University, U.S.A.*
[2]*Flexday Solutions LLC, U.S.A.*
liuj1@ohio.edu

Keywords: Document Analysis, Key Information Extraction (KIE), Graph Neural Network (GNN), BERT.

Abstract: Key Information Extraction (KIE) is a critical and often final step in the comprehensive process of document analysis. Various graph-based solutions, including SDMG-R, have been proposed to address the challenges posed by the relationships between document components. In this paper, we propose a spatial structure-guided framework to integrate known structures of the data and tasks, which are represented as ground-truth graphs. This integration is enforce by minimizing a (dis-)similarity loss defined on graph edges. To optimize graph similarity, different loss functions are explored for the edge loss. In addition, we enhance the text feature extraction component in SDMG-R from character-level Bi-LSTM to word-level embeddings using a fine-tuned BERT, thereby integrating deeper language knowledge into the text labeling procedure. Experiments on the FUNSD and WildReceipt datasets demonstrate the effectiveness of our proposed model in extracting key information from document images with unseen templates, significantly outperforming baseline models.

1 INTRODUCTION

Extracting and understanding meaningful information from structured or unstructured documents plays a crucial role in various applications, including legal document review, archiving and digitization, sentiment and content analysis, and document classification. Key Information Extraction (KIE) from Visually Rich Documents (VRD) is a critical, and often the final, step in document understanding.

While significant strides have been made over the past decade, largely due to the application of deep learning techniques, challenges remain in document analysis and KIE. The necessary features for understanding documents, which include textual, visual, and spatial elements, often exhibit great variation even within the same document type. In addition, the relationships between document components, which can span different visual, semantic, and spatial dimensions, are often difficult to fully comprehend.

To address these two challenges, various graph-based solutions have been proposed (Yu et al., 2021; Liu et al., 2019; Wang et al., 2023; Hong et al., 2022; Qian et al., 2018; Sun et al., 2021; Hwang et al., 2020; Chen et al., 2023; Shi et al., 2023; Gbada et al., 2024b; Lee et al., 2021), achieving state-of-the-art performance. PICK (Yu et al., 2021) incorporates all features from the document—textual, visual, and spatial—where an encoder generates visual embeddings using Convolutional Neural Networks (CNN) and text embeddings using Transformers. These features are fed into a graph module to capture the latent relationships between nodes. For KIE, sequence tagging is performed at the character level using a *Bidirectional LSTM-Conditional Random Field* layer. Inspired by SPADE (Hwang et al., 2020), Hong et al. (Hong et al., 2022) introduced a graph-based method called BROS, which encodes the relative positions of text blocks in a document's 2D space to capture their relationships. BROS has a similar structure to LayoutLM (Xu et al., 2020) but includes two major improvements.

Spatial Dual-Modality Graph Reasoning (SDMG-R) (Sun et al., 2021) attempts to address both challenges. It models document images as dual-modality graphs, with nodes encoding both the visual and textual features of detected text regions, and edges representing the spatial relations between neighboring text regions. However, for data with well-defined structures, such as driver's licenses, passports, and transcripts, their structures have not been fully utilized in the key classification procedure. In addition,

*Corresponding Author. This project is in part supported by Flexday AI.

features extracted from visual and textual modalities have been processed through U-Net and Bi-LSTM, respectively, before being fused and fed into the graph network. The textual feature extraction, however, is conducted at the character level, lacking an understanding of the word content.

In this paper, we propose two novel components to improve the graph-based feature extraction and structure learning procedure. Built on SDMG-R, our framework replaces the character-level network with word-embedding networks, allowing the system to recognize words for text labeling. For structure learning, we integrate the available ground-truth text layout and connections into the graph learning process. Our contributions can be summarized as follows:

- First, we enhance the text feature extraction component in SDMG-R by replacing the character-level Bi-LSTM with a word-level embedding solution based on a fine-tuned BERT model. This improvement incorporates language knowledge directly into the text labeling process.
- Second, we propose the use of a ground-truth graph as both an input and a guiding mechanism. This novel approach sets our method apart and enhances the performance of the baseline SDMG-R. The guidance is enforced through a loss function designed to minimize graph dissimilarity.
- For the graph dissimilarity setup, we experimented with various loss functions, including Mean Squared Error (MSE), Cross-Entropy (CE), and Focal Loss, to evaluate their effectiveness in graph-based learning tasks.

2 TECHNICAL BACKGROUND

Optical Character Recognition **(OCR) based document analysis** is a powerful approach to extracting and processing textual information from images and documents. This technology combines computer vision and machine learning techniques to transform physical or digital documents into machine-readable text, enabling automated information retrieval and analysis (Subramani et al., 2020). The process typically involves three main stages: *text detection*, *text recognition*, and *key information extraction*, each employing networks to handle various challenges such as diverse document layouts, different fonts and text styles, and complex backgrounds.

The first stage, text detection, identifies and localizes text regions within an image or document and the major solutions include (Liao et al., 2020; Wang et al., 2019b; Wang et al., 2019a; Zhang et al., 2020; Zhu et al., 2021; Long et al., 2018). Text recognition then converts these detected regions into actual text, interpreting the visual patterns of characters and words (Fang et al., 2021; Shi et al., 2018; Li et al., 2019; Shi et al., 2016; Sheng et al., 2019; Du et al., 2022). Finally, key information extraction aims to identify and extract specific, relevant data from the recognized text (Yu et al., 2021; Sun et al., 2021). This three-stage pipeline to end-to-end document analysis has wide-ranging applications across industries, from digitizing historical archives to automating data entry in business processes.

KIE faces many challenges due to the variations in document types, feature diversity, and the complex relationships between different document components (Gbada et al., 2024a). To address these challenges, two primary strategies have been employed: graph-based and end-to-end approaches. Graph-based solutions (Yu et al., 2021; Liu et al., 2019; Krieger et al., 2021; Wang et al., 2023) model the relationships between document components, using nodes and edges to represent different features or relationships. End-to-end approaches (Xu et al., 2020; Katti et al., 2018; Shehzadi et al., 2024), on the other hand, integrate the entire extraction process into a single, cohesive framework. Graph-based methods, in particular, have proven to be effective in capturing dependencies among document entities (Qian et al., 2018; Sun et al., 2021; Hwang et al., 2020; Xu et al., 2020; Hong et al., 2022).

3 METHOD

The baseline model in this work is SDMG-R, as shown in Fig. 1. SDMG-R begins with the identification of bounding boxes within the document, which serve as the primary units for classification. A multimodal framework is employed, leveraging both visual and textual information from these bounding boxes to capture essential features. After fusing this multimodal data, a graph is constructed where each node represents a fused feature. The edges, representing intrinsic relationships such as item-price pairs, encapsulate critical information for classification. Both node and edge information are integrated to achieve accurate results.

A two-pronged strategy is employed for feature extraction, leveraging advanced techniques in both visual and textual domains. Visual features are extracted using a pretrained U-Net (Ronneberger et al., 2015), known for its strong performance in image segmentation. Bounding boxes are processed, followed by Region of Interest (ROI) pooling to

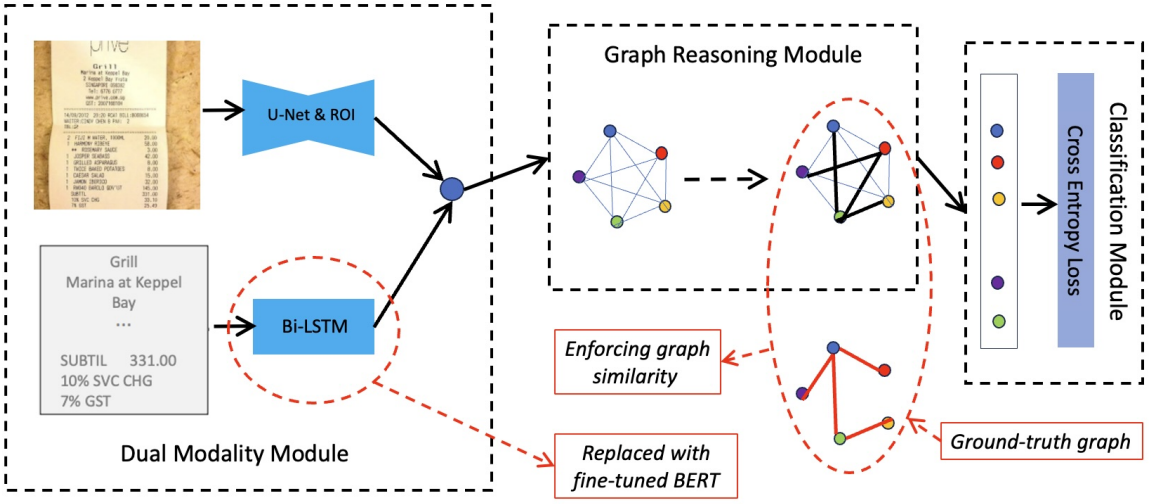

Figure 1: Baseline SDMG-R model and our proposed components. The proposed components are highlighted with red color, which include 1) relacing Bi-LSTM with BERT; and 2) a new loss to enforce the similarity between the learned graph and ground-truth graph.

standardize varying sizes. For textual features, a character-based Bi-LSTM network is trained in conjunction with the entire model to effectively capture textual context.

Despite the careful design and performance improvements in KIE achieved through feature extraction and graph reasoning steps, both approaches have significant limitations. The feature extraction network, trained on local data at the character level, fails to leverage pre-trained word embeddings from large text datasets, potentially missing out on deeper semantic understanding of words. Additionally, the key classification procedure does not fully exploit the well-defined structures inherent in documents like driver's licenses, passports, and transcripts. These standardized formats could provide valuable cues for more accurate information extraction, yet remain underutilized in current methodologies.

To address these limitations, we propose two key components, highlighted in red in Fig. 1: (1) replacing Bi-LSTM with a fine-tuned BERT to enhance text feature extraction, and (2) introducing a new loss function to enforce similarity between the learned graph and the ground-truth graph. In addition, we explore different choices for the loss function to enforce graph similarity. Detailed explanations of these components will be presented in the following subsections.

3.1 Textual Feature Extraction

The textual feature extraction component in SDMG-R processes text using a Bi-LSTM at the character level. This component is trained with local data to predict next characters, without grasping the semantic meaning of the words. This approach is akin to recognizing patterns in a foreign language based solely on their occurrence in training data, without comprehending their meaning. As shown in Fig. 2, the BI-LSTM might be able to identify "grill" in a document as a five-character sequence, "g", "r", "i", "l", and "l", but fail to associate it with a restaurant name. This character-level processing limits the system's ability to leverage contextual clues that could enhance information extraction.

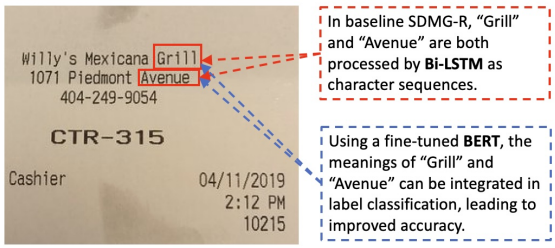

Figure 2: Illustration of how character-level Bi-LSTM and a fine-tuned BERT would process words on a receipts differently.

To overcome the limitations of the baseline model, we propose integrating pre-trained word embeddings like word2vec (Mikolov, 2013), GloVe (Pennington et al., 2014), and BERT (Devlin, 2018). These advanced language models capture word meanings and relationships, offering a powerful tool to improve semantic comprehension. By incorporating these pre-trained embeddings, we can significantly boost the system's ability to accurately classify documents and

extract key information, especially when dealing with domain-specific terminology or contextual subtleties. This is illustrated in Fig. 2, where a word embedding model recognizes the meanings of "Grill" and "Avenue" increasing the likelihood of correct classification into their labels as *restaurant name* and *address*.

In this paper, we fine-tune a pre-trained BERT model on a token classification task with a dataset labeled with *Begin-Inside-Outside* tagged (BIO-tagged) tokens. The fine-tuning process employs tokenized inputs generated by BERT's tokenizer and utilizes its pre-trained contextual embeddings to capture the context of each token. Through this task-specific training, BERT learns to accurately assign BIO labels, making it highly effective for token classification tasks. The resulting fine-tuned model is then used in this work to generate refined word embeddings.

Another benefit of using BERT is its robustness against errors from the previous text recognition step. For instance, if the OCR step incorrectly recognizes "Grill" in Fig. 2 as "Griil", BERT can potentially correct this error, still achieving accurate recognition. This provides a significant advantage over the baseline Bi-LSTM for inputs with poor image quality.

3.2 Enforcing Structure with Ground-Truth Graphs

In the baseline SDMG-R's graph reasoning module, as shown in Fig. 1, the model uses a graph network to implicitly learn the connectivity among document items (graph nodes). The graphs are formulated as fully connected graphs, and the weights between nodes, which reflect relationships, are learned to help label the nodes correctly. However, ground-truth connectivities are not explicitly enforced. In other words, the model neither rewards correct or relevant connections nor penalizes incorrect or irrelevant connections learned within its graph network.

This poses a significant limitation, as many well-structured documents have clearly defined connections among their items. For instance, Fig. 3 illustrates an example using a section of a U.S. passport. The spatial relationships and connectivity between the "Nationality" field and the text blocks "UNITED", "STATES", "OF", and "AMERICA" are consistent across all U.S. passports. Such ground-truth connectivity is often available in the training data. Failing to take use of these connections would be a waste of valuable information.

To address this limitation, we propose a graph (dis)similarity loss that enforces the similarity between the learned graphs and the ground-truth structures. These ground-truth graphs are constructed

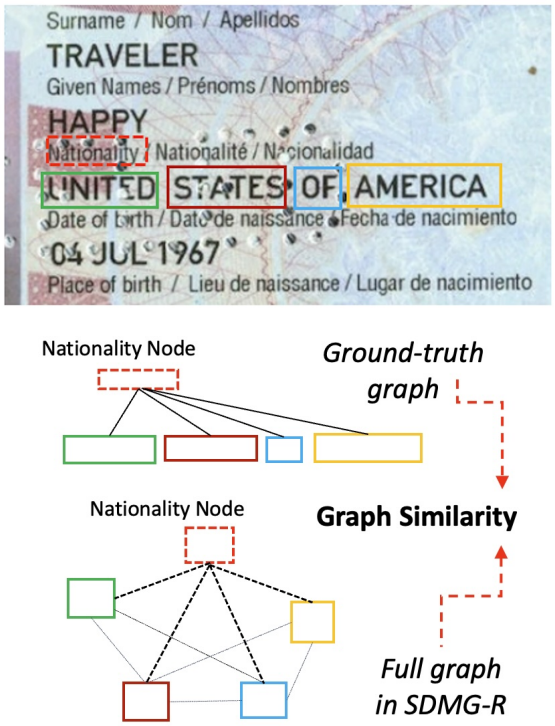

Figure 3: Illustration of our proposed graph similarity component. Top: An example from U.S. passports, where the "Nationality" node is connected to the four items below. Bottom: This ground-truth structure can be enforced in SDMG-R by incorporating a graph (dis-)similarity loss.

based on key-value pairs within the documents, with each edge indicating whether two nodes are truly connected (true) or not (false). The graph similarity task is framed as an edge classification problem, where the predicted edge weights are driven to align closely with the ground-truth values. Minimizing this loss encourages the graph reasoning module in SDMG-R to generate connections that accurately reflect the document structure, thereby improving node classification accuracy and aligning it with the actual relationships present in the document layout.

In Fig. 3, this means that we encourage SDMG-R to predict that the "Nationality" node is connected to the other four nodes while ensuring that there are no connections between the four nodes themselves. Achieving this would be beneficial for accurately labeling all five nodes.

3.2.1 Different Setups for Graph (dis-)Similarity Loss

As our graph similarity component is formulated as a binary edge classification task, the choice of loss function is crucial and can be set up in various ways. The following options have been explored in this

work.

Cross entropy (CE) loss or log loss is commonly used to measure the performance of a model. It quantifies the difference between the predicted probabilities and the actual binary labels (0 or 1). The cross-entropy loss function is defined as:

$$\text{Loss} = -(y\log(p) + (1-y)\log(1-p))$$

Here, y is the actual label (0 or 1), and p is the predicted probability that the label is 1.

Focal loss adds a modulating term $(1-p_t)^\gamma$ to cross-entropy loss to account for class imbalance during training tasks like object detection. The formula is given in (Lin, 2017) as:

$$FL(p_t) = -(1-p_t)^\gamma \log(p_t)$$

Setting $\gamma > 0$ can reduce focus on well-classified examples and put more emphasis on misclassified examples.

MSE loss measures the average squared difference between the predicted values and actual values. The formula is given as:

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$$

Here, m is the number of samples, $y^{(i)}$ denotes the ground truth label for the i-th sample, and $\hat{y}^{(i)}$ denotes the predicted label for the i-th sample.

4 EXPERIMENTS AND RESULTS

In this section, we present and evaluate the experimental results of the proposed models. The competing model were tested on two different datasets: 1) the public *Form Understanding in Noisy Scanned Documents* (FUNSD) dataset (Jaume et al., 2019), and 2) the WildReceipt dataset, introduced in (Sun et al., 2021). To enhance textual feature extraction, a uncased BERT based model downloaded from `https://huggingface.co/google-bert/bert-base-uncased` was fine-tuned on both the FUNSD and WildReceipt datasets for the token classification task using BIO tagging.

4.1 Results from FUNSD Dataset

The FUNSD dataset consists of 199 fully annotated scanned form images, specifically designed for text detection, OCR, and document understanding tasks (Jaume et al., 2019). However, the forms are noisy and vary significantly in appearance, making it a difficult dataset for document understanding (Jaume et al., 2019). The dataset includes 149 training images and 50 testing images, with each form image accompanied by a JSON file containing its annotations. Each form includes a list of interlinked semantic entities, where a semantic entity represents a group of words, each assigned a unique ID, label, bounding box coordinates, and relationships with other words. These relationships define the key-value pairs using ID numbers. The nodes in FUNSD have four labels: *header, question, answer*, and *other*. For our performance comparison, we exclude the *other* class. The ground-truth graphs for FUNSD samples are generated using the relationships defined in the dataset.

Table 1 presents the results of the baseline model and our proposed models on the FUNSD dataset, using various edge losses and textual feature extractors. All our models outperform the baseline SDGM-R, which does not have the edge-loss component. Among our models, those using fine-tuned BERT outperform their counterparts using Bi-LSTM, highlighting BERT's superiority as a feature extractor. Focal loss also consistently outperforms MSE and CE across both BERT and Bi-LSTM feature extractors. The final combination, fine-tuned BERT with focal loss, achieves the highest macro F1-score of 83.41%.

Table 1: Results from the segmentation models with different edge losses and textual feature extractors on the FUNSD dataset. "FT BERT" stands for fine-tuned BERT.

Textual Feature	Edge Loss	Macro F-1 Score
Baseline (SDMG-R)		75.66
Bi-LSTM	*MSE*	77.45
	Cross Entropy	77.82
	Focal Loss	78.43
Fine-tuned BERT	*MSE*	81.22
	Cross Entropy	81.55
	Focal Loss	**83.41**

Fig. 4 illustrates an example of a FUNSD form where the baseline model misclassifies certain text, while our approach delivers accurate classification. In this figure, *header* classes are represented by blue rectangles, *questions* by red rectangles, *answers* by green rectangles, and *other* classes by orange rectangles. From Fig. 4(a) and 4(b), we observe that two *other* classes are incorrectly classified as *header* classes by the original SDMG-R model. In addition, one *header* class is misclassified as a *question* class. However, in Fig. 4, all these misclassifications are corrected.

In the end, we compared our proposed model with recent studies (Lee et al., 2021; Hwang et al., 2020; Gbada et al., 2024b; Chen et al., 2023; Shi et al., 2023; Hong et al., 2022) that reported results on the FUNSD dataset for the KIE task, summarized in

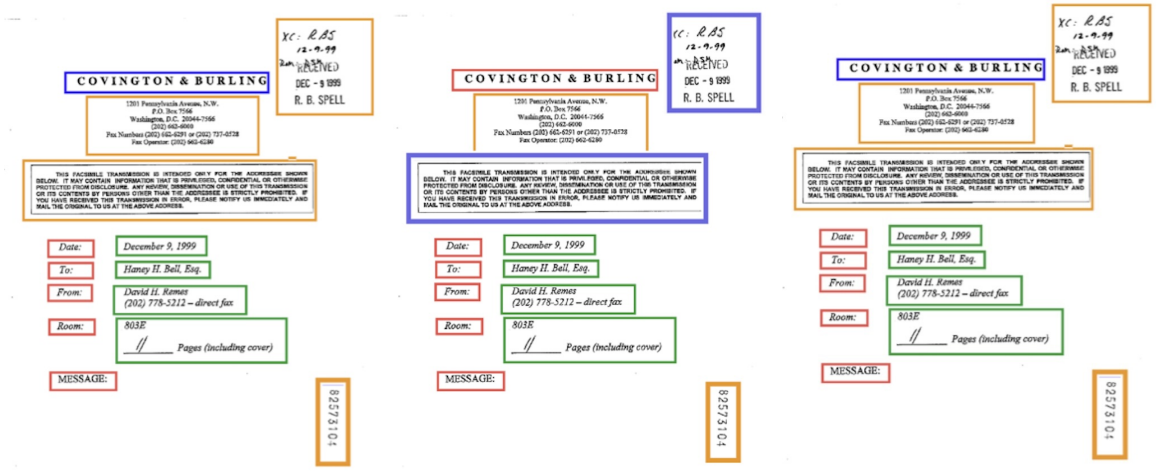

Figure 4: Classification results on FUNSD forms. Colors indicate different classes: blue for *header*, red for *questions*, green for *answers*, and orange for *other* classes. Left: Ground-truth classes of the text blocks. Middle: Predicted classes using original SDMG-R. Right: Predicted classes using our proposed model.

(Gbada et al., 2024a). It is important to note that we do not intend to make a direct quantitative comparison among these methods, as the studies employed different training setups and evaluation metrics. Nevertheless, the classification accuracy achieved by our model (fine-tuned BERT with focal loss) is comparable to that of the best-performing reported model.

Table 2: Results from different models on FUNSD dataset for the KIE task. The metric used in results is the F1-Score.

Related Work	Results
ROPE (Lee et al., 2021)	57.22
SPADE (Hwang et al., 2020)	72
Gbada et al. (Gbada et al., 2024b)	80.4
DAMGCN (Chen et al., 2023)	80.63
LayoutGCN (Shi et al., 2023)	82.06
BROS (Hong et al., 2022)	**84.52**
Our Proposed Model	**83.41**

4.2 Results from WildReceipt Dataset

The WildReceipt dataset, introduced in (Sun et al., 2021), is designed for OCR and KIE tasks. It includes a collection of 1,765 receipt images, divided into 1,267 training images and 472 testing images. Each image contains a list of OCR entries, with each entry comprising a bounding box, text, and a class label. The dataset defines 26 classes, which include various key-value pairs such as *Store_name_key* vs *Store_name_value*, *Date_key* vs *Date_value*, and *Total_key* vs *Total_value*.

The original WildReceipt dataset does not contain relationships between nodes. In this work, we introduced relationships between key-value pairs, i.e., connecting *name_keys* with *name_values*, and take the graphs as the ground-truth in our models. Following the same practice in the baseline model (Sun et al., 2021), we excluded the scores for the *Ignore* and *Others* classes from the performance evaluation.

Table 3 shows the KIE results from the baseline SDMG-R model and our models. All of our models outperform the baseline, which utilizes a Bi-LSTM and lacks edge loss. Among our six models, a similar trend to the FUNSD experiments can be observed. Regarding the text feature extractor, the three models with fine-tuned BERT consistently outperform the corresponding models using Bi-LSTM. For edge loss, focal loss enhances accuracy more effectively than MSE or cross-entropy. This can be attributed to the fact that the ground-truth graphs are typically quite sparse in edges, making focal loss more suitable for such label-imbalanced scenarios.

Table 3: Results from the segmentation models with different edge losses and textual feature extractors on the WildReceipt dataset.

Textual Feature	Edge Loss	Macro F-1 Score
Baseline (SDMG-R)		88.35
Bi-LSTM	*MSE*	88.91
	Cross Entropy	88.40
	Focal Loss	89.20
Fine-tuned BERT	*MSE*	89.85
	Cross Entropy	89.72
	Focal Loss	**90.26**

To summarize, our two design components—replacing Bi-LSTM with fine-tuned BERT and enforcing graph structure using an edge loss—successfully achieved the intended design goals, as clearly demonstrated in both the FUNSD

and WildReceipt experiments. Among the three loss functions explored (MSE, cross-entropy, and focal loss), focal loss consistently delivered the best performance across datasets and experiments. This can be largely attributed to the sparsity of the ground-truth graphs, where focal loss is better suited to handle label imbalance.

5 CONCLUSIONS

In this paper, we present a spatial structure-guided framework to address the challenges of KIE by leveraging ground-truth graphs and optimizing graph similarity through various loss functions. Moreover, by enhancing the text feature extraction process with word-level embeddings using a fine-tuned BERT, our models demonstrate superior performance compared to the baseline SDMG-R model. Experimental results on the FUNSD and WildReceipt datasets confirm the robustness of our approach.

REFERENCES

Chen, Y.-M., Hou, X.-T., Lou, D.-F., Liao, Z.-L., and Liu, C.-L. (2023). Damgcn: Entity linking in visually rich documents with dependency-aware multimodal graph convolutional network. In *International Conference on Document Analysis and Recognition*, pages 33–47. Springer.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., and Jiang, Y.-G. (2022). Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*.

Fang, S., Xie, H., Wang, Y., Mao, Z., and Zhang, Y. (2021). Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7098–7107.

Gbada, H., Kalti, K., and Mahjoub, M. A. (2024a). Deep learning approaches for information extraction from visually rich documents: datasets, challenges and methods. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–22.

Gbada, H., Kalti, K., and Mahjoub, M. A. (2024b). Multimodal weighted graph representation for information extraction from visually rich documents. *Neurocomputing*, 573:127223.

Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., and Park, S. (2022). Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10767–10775.

Hwang, W., Yim, J., Park, S., Yang, S., and Seo, M. (2020). Spatial dependency parsing for semi-structured document information extraction. *arXiv preprint arXiv:2005.00642*.

Jaume, G., Ekenel, H. K., and Thiran, J.-P. (2019). Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., and Faddoul, J. B. (2018). Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.

Krieger, F., Drews, P., Funk, B., and Wobbe, T. (2021). Information extraction from invoices: a graph neural network approach for datasets with high layout variety. In *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*, pages 5–20. Springer.

Lee, C.-Y., Li, C.-L., Wang, C., Wang, R., Fujii, Y., Qin, S., Popat, A., and Pfister, T. (2021). Rope: reading order equivariant positional encoding for graph-based document information extraction. *arXiv preprint arXiv:2106.10786*.

Li, H., Wang, P., Shen, C., and Zhang, G. (2019). Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8610–8617.

Liao, M., Wan, Z., Yao, C., Chen, K., and Bai, X. (2020). Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481.

Lin, T. (2017). Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Liu, X., Gao, F., Zhang, Q., and Zhao, H. (2019). Graph convolution for multimodal information extraction from visually rich documents. *arXiv preprint arXiv:1903.11279*.

Long, S., Ruan, J., Zhang, W., He, X., Wu, W., and Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36.

Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Qian, Y., Santus, E., Jin, Z., Guo, J., and Barzilay, R. (2018). Graphie: A graph-based framework for information extraction. *arXiv preprint arXiv:1810.13083*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Shehzadi, T., Stricker, D., and Afzal, M. Z. (2024). A hybrid approach for document layout analysis in document images. In *International Conference on Document Analysis and Recognition*, pages 21–39. Springer.

Sheng, F., Chen, Z., and Xu, B. (2019). Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 781–786. IEEE.

Shi, B., Bai, X., and Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X. (2018). Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048.

Shi, D., Liu, S., Du, J., and Zhu, H. (2023). Layoutgcn: A lightweight architecture for visually rich document understanding. In *International Conference on Document Analysis and Recognition*, pages 149–165. Springer.

Subramani, N., Matton, A., Greaves, M., and Lam, A. (2020). A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*.

Sun, H., Kuang, Z., Yue, X., Lin, C., and Zhang, W. (2021). Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*.

Wang, J., Krumdick, M., Tong, B., Halim, H., Sokolov, M., Barda, V., Vendryes, D., and Tanner, C. (2023). A graphical approach to document layout analysis. In *International Conference on Document Analysis and Recognition*, pages 53–69. Springer.

Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., and Shao, S. (2019a). Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9336–9345.

Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., and Shen, C. (2019b). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8440–8449.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020). 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*, pages 1192–1200.

Yu, W., Lu, N., Qi, X., Gong, P., and Xiao, R. (2021). Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. IEEE.

Zhang, S.-X., Zhu, X., Hou, J.-B., Liu, C., Yang, C., Wang, H., and Yin, X.-C. (2020). Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9699–9708.

Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., and Zhang, W. (2021). Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131.