Doc2Graph-X: A Multilingual Graph-Based Framework for Form Understanding

Souparni Mazumder[§⋆], Sanket Biswas[§1], Alloy Das[2], and Josep Lladós[1]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Catalonia, Spain
raim8218@gmail.com, {sbiswas, josep}@cvc.uab.es
[2] Habitat Labs, Habitat Lens Private Limited, Kolkata, India
alloy@habitatlens.in

Abstract. Graph Neural Networks (GNNs) have advanced key information extraction (KIE) in document AI, but existing methods remain monolingual. We introduce Doc2Graph-X, a multilingual extension of Doc2Graph, leveraging word-level and sentence-level embeddings for robust cross-lingual document representation. Our framework constructs graph-based structures where a node classifier performs semantic entity recognition (SER) and an edge classifier handles relation extraction (RE) to predict links between entities. Evaluated on the XFUND dataset across seven languages, Doc2Graph-X outperforms existing baselines, demonstrating strong multilingual adaptability. Additional results on FUNSD validate its effectiveness in monolingual settings. Our approach enables structured multilingual document understanding while preserving task-agnostic adaptability. The code[†] will be made available upon acceptance.

Keywords: Multilingual Document Understanding · Graph Neural Networks · Key Information Extraction · Entity Detection and Linking

1 Introduction

Understanding documents is a fundamental challenge in Document AI [6], particularly in structured formats like forms, invoices, and reports, where extracting key information and analyzing relationships between elements (e.g., titles, structured text, footnotes, tables, figures etc.) is crucial. Traditionally, Graph Neural Networks (GNNs) [14,19,12] have been widely adopted for document structure representation [11,10], effectively modeling relationships between textual and visual components. However, most existing GNN-based methods [20,11,10] operate in monolingual settings, limiting their effectiveness in a world where multilingual document processing is increasingly essential across industries such as finance, healthcare, and legal technology.

⋆ Work done during internship at Computer Vision Center
§ Equal contribution
† https://github.com/biswassanket/doc2graph_multi.git

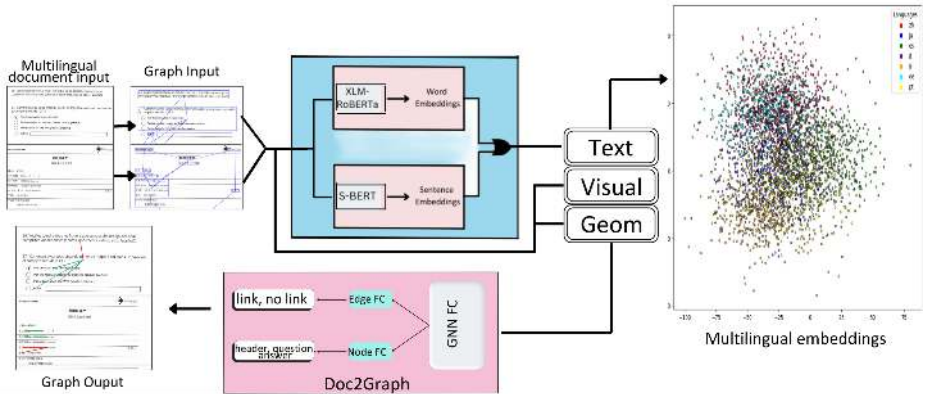

Fig. 1: **Overview of the Doc2Graph-X Framework.** Multilingual documents are converted into graph representations using either word-level or sentence-level embeddings. The GNN processes textual, visual, and geometric features, with a node classifier for SER and an edge classifier for RE task, enabling structured multilingual document understanding.

To address this gap, we introduce `Doc2Graph-X`, a multilingual extension of Doc2Graph [11], designed to enable language-agnostic structured document understanding. Unlike its predecessor, which relied on monolingual text processing pipelines [15], Doc2Graph-X utilizes Sentence-BERT [17] for sentence-level embeddings and XLM-RoBERTa [5] for word-level embeddings, enabling multilingual document processing. These embeddings serve as the foundation for constructing a graph-based document representation, where nodes correspond to semantic entities such as those commonly found in forms (e.g., names, dates, addresses, amounts, and table entries), and edges encode learned relationships between them. A multimodal GNN [11,12] processes textual, visual, and geometric features simultaneously to refine these connections, allowing it to distinguish entity roles and predict meaningful relationships as illustrated in Figure 1. The node classifier performs Semantic Entity Recognition (SER) by detecting key entities, while the edge classifier handles Relation Extraction (RE) by predicting links between entities as binary relations (0 or 1). This structured learning process enables Doc2Graph-X to generalize across multiple languages while maintaining interpretability in complex document layouts.

In this work, we establish Doc2Graph-X as a graph-based multilingual document understanding framework, leveraging graph learning to model complex relationships within multilingual forms. Our approach makes three key contributions: (1) Multilingual Graph Representations, where SBERT and XLM-RoBERTa embeddings enable language-agnostic entity detection and linking; (2) Multimodal Graph Learning, where a Graph Neural Network (GNN) jointly

processes textual, visual, and geometric features to capture structured document relationships; and (3) Graph-Based Benchmarking on XFUND [23], where we introduce the first graph-based evaluation for multilingual forms, demonstrating superior performance over existing baselines.

2 Related Work

GNNs were initially introduced for Document Layout Analysis (DLA) tasks [2], such as table detection [18] and table structure recognition [16]. The motivation behind their adoption was two-fold: (i) leveraging the geometric properties of documents for structural analysis and (ii) preserving the privacy of textual content, particularly in administrative documents, by making models language-independent and structure-driven [18,1]. For instance, GNN-based table detection in invoices [18] demonstrated that documents could be effectively processed without relying on raw text. Carbonell et al. [3] extended this idea by employing relational Graph Convolutional Networks (GCNs) [14] for form analysis in FUNSD benchmark [13], tackling word-level entity grouping, labeling, and linking. Their approach relied on bounding box information and word embeddings as node features while using k-nearest neighbors (KNNs) to define edge connections. However, this method lacked visual features, which limited its ability to capture layout-based relationships. Building on this, the FUDGE framework [10] integrated relationship pairs using a lightweight CNN-based detection model [9] and subsequently applying a GCN, FUDGE enriched text entity classification and key-value pair extraction. Doc2Graph [11] introduced an end-to-end task-agnostic GNN approach to structured document understanding for node classification (SER task) and edge classification (RE task), demonstrating superior performance while maintaining the most optimal number of trainable parameters. However, a major limitation of Doc2Graph was its monolingual nature, restricting its applicability to multilingual document processing.

To address multilingual form understanding, the XFUND benchmark [23] was introduced, providing a large-scale multilingual dataset spanning seven languages. Alongside XFUND, LayoutXLM [22], LiLT [21], and Graph-MLLM [7] were proposed as a transformer-based multilingual baseline, setting an initial benchmark for multilingual key information extraction (KIE) in visually rich documents. In this work, we build on Doc2Graph and extend it to Doc2Graph-X, introducing multilingual embeddings and multimodal graph learning. By benchmarking Doc2Graph-X on XFUND, we explore the potential of GNN-based models in the challenging multilingual form understanding setting. Our approach integrates sentence-level [17] and word-level embeddings [5], leveraging graph-based reasoning to capture structured relationships across languages while maintaining task-agnostic adaptability.

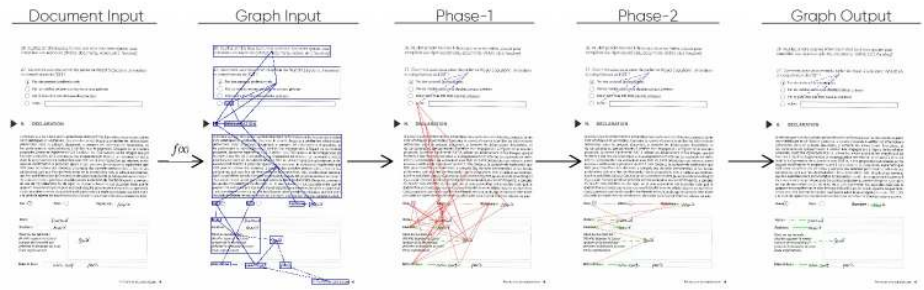

Fig. 2: **Graph-based document processing in Doc2Graph-X.** The document is first converted into a fully connected graph using the K-NN algorithm, linking all entities. A multimodal GNN then applies a message passing algorithm in two phases: Phase 1 refines initial connections by learning entity relationships, and Phase 2 further prunes irrelevant edges while strengthening key links. The final graph output retains only meaningful relationships, enabling accurate SER and RE for structured document understanding.

3 Method

We introduce `Doc2Graph-X`, an extension of Doc2Graph, designed to process multilingual documents using multimodal graph learning. The overall approach has been designed to integrate textual, visual, and spatial features into a graph-based framework, enabling language-agnostic entity recognition and relationship extraction while preserving document structure and layout.

Graph Representation of Documents. A document is represented as an undirected graph $G = (V, E)$, where V is the set of nodes representing document entities (e.g., words, sentences, tables), and E is the set of edges capturing relationships between entities. Each node $v_i \in V$ is associated with a feature vector $\mathbf{h}_i$ that encodes information from multiple modalities:

$$\mathbf{h}_i = [\mathbf{t}_i; \mathbf{v}_i; \mathbf{l}_i] \tag{1}$$

where: $\mathbf{t}_i$ is the textual embedding from `SBERT` at sentence-level or `XLM-RoBERTa` at word-level, $\mathbf{v}_i$ is the visual feature extracted using `U-Net`, $\mathbf{l}_i$ is the layout feature capturing spatial position. The edges $e_{ij} \in E$ define the relationships between entities and are weighted based on learned connections. The edge features are given by:

$$\mathbf{e}_{ij} = [d_{ij}; \theta_{ij}] \tag{2}$$

where d_{ij} is the normalized Euclidean distance and θ_{ij} is the relative angle in polar coordinates. Unlike Doc2Graph [11], which was monolingual, `Doc2Graph-X` enables language-agnostic document processing by leveraging the cross-lingual embeddings, without requiring task-specific models for each language.

Graph-based Document Processing. To ensure local connectivity as shown in graph input of Figure 2, edges are initialized using the `K-Nearest Neighbors (KNN)` algorithm, where each node $v_i \in V$ is connected to its K-nearest neighbors based on Euclidean distance:

$$e_{ij} = \exp\left(-\frac{d(v_i, v_j)^2}{2\sigma^2}\right) \tag{3}$$

where $d(v_i, v_j)$ is the Euclidean distance between node positions, and σ is a scaling factor. This step maintains document structure while preventing over-connectivity.

Learning based Edge Refinement. As shown in Phase 1 of Figure 2 the node embeddings are then updated using a message-passing GNN:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(W \cdot \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{h}_j^{(l)}\right) \tag{4}$$

where $\mathcal{N}(i)$ is the neighborhood of node i, α_{ij} is an attention weight for neighboring nodes, W is a learnable weight matrix, σ is an activation function. The edges are refined using an edge classifier:

$$\hat{r}_{ij} = \sigma\left(W_e \cdot [\mathbf{h}_i; \mathbf{h}_j; d_{ij}; \theta_{ij}]\right) \tag{5}$$

where W_e is a learnable weight matrix, d_{ij} is the Euclidean distance, and θ_{ij} is the angular displacement in polar coordinates. Edges with $\hat{r}_{ij} > \tau$ (where τ is a learned threshold) are retained, while weaker edges are removed.

Graph Learning Objectives. The model jointly optimizes two tasks: Semantic Entity Recognition (SER) and Relation Extraction (RE). The objective function is defined as:

$$\mathcal{L} = \sum_{i \in V} \mathcal{L}_{\text{SER}}(y_i, \hat{y}_i) + \sum_{(i,j) \in E} \mathcal{L}_{\text{RE}}(r_{ij}, \hat{r}_{ij}) \tag{6}$$

where: $\mathcal{L}_{\text{SER}}(y_i, \hat{y}_i)$ is the node classification loss for SER, $\mathcal{L}_{\text{RE}}(r_{ij}, \hat{r}_{ij})$ is the edge classification loss for RE. Each term represents a standard cross-entropy loss, ensuring accurate node and edge predictions. The final loss function balances both tasks.

4 Experimental Validation

In this section, we present the quantitative and qualitative evaluation of Doc2Graph-X on Semantic Entity Recognition (SER) and Relation Extraction (RE) tasks. The model is benchmarked against transformer-based methods to highlight the effectiveness of graph-based learning with significantly fewer parameters.

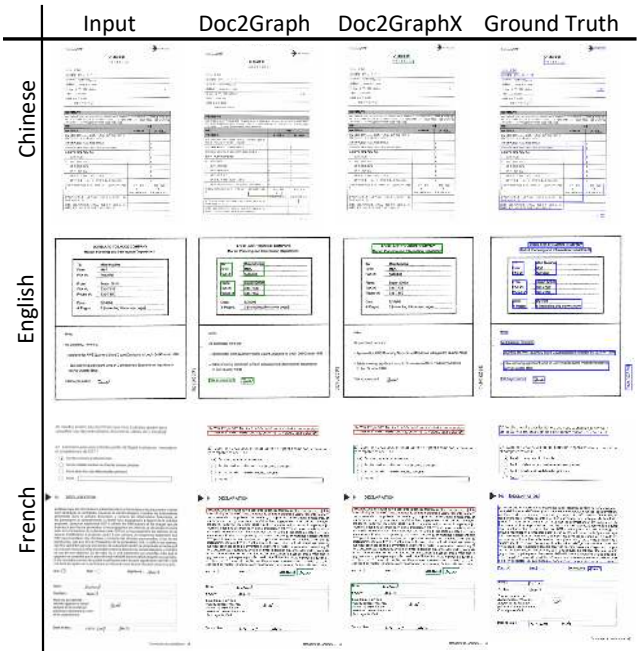

Fig. 3: **Qualitative comparison of SER across multiple languages** The figure compares Doc2Graph-X with Doc2Graph and ground truth annotations on Chinese, English, and French documents, demonstrating improved entity detection and alignment in multilingual forms.

Datasets and Evaluation Metrics. The evaluation is conducted on two widely used structured document understanding benchmarks: `FUNSD` [13] and `XFUND` [23]. FUNSD consists of English-scanned forms annotated for key-value entity linking. XFUND extends FUNSD to a multilingual setting, covering seven additional languages: Chinese, Japanese, French, German, Italian, Spanish, and Portuguese. Both datasets provide entity-level annotations and relationship links, making them suitable for evaluating *multilingual document parsing and entity linking*. The model performance is assessed using the *micro F1 score*, which balances precision and recall in entity recognition and relation prediction tasks. For SER, the F1 score measures the accuracy of entity detection and classification. For RE, it evaluates the correctness of predicted entity relationships.

Quantitative Analysis. Table 1 presents the F1 scores for SER and RE across multiple languages. Despite having only *6.2M parameters*, Doc2Graph-X achieves performance comparable to state-of-the-art transformer-based models, such as LayoutXLM (345M parameters) and InfoXLM (362M parameters).

For SER, Doc2Graph-X achieves an F1 score of 80.07 on FUNSD, surpassing the 79.40 of LayoutXLM [22]. The model generalizes well to multiple languages, with particularly showing very strong performance in French (79.09) and Ital-

Table 1: **Multi Language fine-tuning F1 score on FUNSD and XFUND** (fine-tuning on 8 languages all, testing on X). "SER" denotes semantic entity recognition and "RE" denotes relation extraction.

	Model	#Params	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.↑
SER	XLM-RoBERT$_{\text{BASE}}$ [5]	125M	66.70	87.74	77.61	61.05	67.43	66.87	68.14	68.18	70.47
	InfoXLM$_{\text{BASE}}$ [4]	362M	68.52	88.68	78.65	62.30	70.15	67.51	70.63	70.08	72.07
	LayoutXLM$_{\text{BASE}}$ [22]	345M	79.40	**89.24**	**79.21**	**75.50**	79.02	80.82	**82.22**	**79.03**	**80.56**
	Doc2Graph-X	**6.2M**	**80.07**	81.32	72.18	71.58	**79.09**	**81.38**	78.02	75.46	77.39
RE	XLM-RoBERT$_{\text{BASE}}$ [5]	125M	26.59	51.05	58.00	52.95	49.65	53.05	50.41	39.82	47.69
	InfoXLM$_{\text{BASE}}$ [4]	362M	29.20	52.14	60.00	55.16	49.13	52.81	52.62	41.70	49.10
	LayoutXLM$_{\text{BASE}}$ [22]	345M	54.83	**70.73**	**69.63**	**68.96**	**63.53**	**64.15**	**65.51**	**57.18**	**64.32**
	Doc2Graph-X	**6.2M**	**54.92**	54.81	48.59	42.28	52.17	55.92	45.94	41.91	49.56

Table 2: **Comparison of F1 scores for SER and RE on FUNSD**.

	F_1 (↑)	
Model	**SER**	**RE**
Doc2Graph [11]	**0.8225** $\pm$ 0.005	0.5336 $\pm$ 0.036
Doc2Graph-X [SBERT]	0.8007	0.5492
Doc2Graph-X [XLM-R]	0.7890	**0.5543**

ian (81.38). For RE, Doc2Graph-X attains an F1 score of 54.92 on FUNSD, surpassing InfoXLM and XLM-RoBERTa while closely beating the much larger LayoutXLM model (54.83). The results indicate that graph-based learning enables effective relation extraction, even with significantly fewer parameters.

Our evaluation on FUNSD (English) and XFUND (multilingual) benchmarks demonstrates the effectiveness of these modifications. While spaCy embeddings [15] retained strong performance in monolingual settings as shown in Table 2, they struggled with cross-lingual generalization. In contrast, Sentence-BERT [17] and XLM-RoBERTa [5] significantly improved performance in multilingual tasks, particularly in semantic entity recognition (SER) and relation extraction (RE), reinforcing the importance of pretrained multilingual embeddings in handling cross-lingual variability.

Qualitative Analysis. Figure 3 and Figure 4 provide qualitative examples of SER and RE predictions respectively.
- For *Semantic Entity Recognition (SER)*, Doc2Graph-X task results are consistent across languages, handling variations in script structure. As depicted in Figure 3, the model preserves document layout integrity, ensuring key entities are recognized correctly even when embedded within complex tabular structures or very dense text regions. There are minor segmentation errors in non-Latin scripts, particularly in Chinese, where closely spaced characters lead to slight misclassification of tokens. The overlapping or dense content areas sometimes lead to fragmented entity detections, though overall detection remains robust.
- *Relation Extraction (RE)* Analysis as depicted in Figure 4 presents the entity

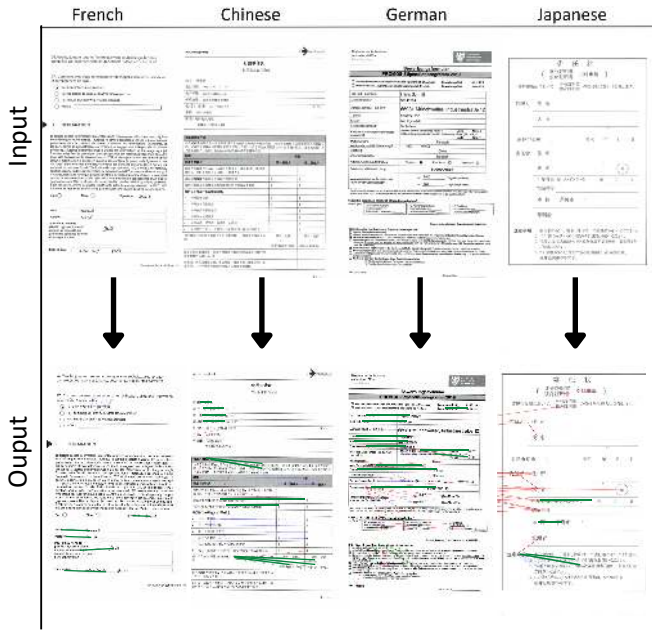

Fig. 4: **Multilingual Entity Linking Results.** The figure shows relation extraction across French, Chinese, German, and Japanese documents. Green lines indicate correct entity links, while red lines denote errors. The results highlight Doc2Graph-X's ability to model structured relationships across languages.

linking (relation extraction) results, where the model predicts semantic relationships between detected entities. We observe that the model correctly identifies (Green edges) key-value pair relationships, even in different writing directions (e.g., left-to-right in French vs. top-to-bottom in Chinese). It can handles nested entities, linking fields correctly within tabular structures, particularly in German forms with very dense metadata sections. There are minimal false positives, ensuring most connections are meaningful and aligned with document semantics. Some missed links (highlighted in Red edges) in Japanese and Chinese forms, are likely due to language-specific variations in sentence structure. The edge ambiguity in complex tables, where relationships are less explicit due to layout variations or missing contextual cues. Despite these challenges, Doc2Graph-X consistently boosts capturing document structure, leveraging graph-based reasoning to improve both entity recognition and relation extraction.

5 Conclusion and Future Work

Doc2Graph-X leverages graph-based learning to effectively preserve spatial and semantic relationships within documents. This approach not only enables robust

multilingual adaptability—ensuring high accuracy in SER and RE tasks—but also reduces false positives in entity linking by explicitly modeling document structure. Remarkably, with just 6.2M parameters, the model delivers competitive performance compared to transformer models that are over 50 times larger, underscoring the efficiency and practicality of GNN-based methods for structured text extraction. Looking ahead, future work will enhance language-specific feature integration to improve cross-lingual adaptability [8], especially in low-resource languages. Exploring hybrid architectures, such as Graphformers [24], promises to unlock synergistic benefits between GNNs and transformers for more nuanced contextual reasoning in complex document structures.

Acknowledgment

This work has been partially supported by the Spanish project PID2021-126808OB-I00 funded by MCIN/AEI/ 10.13039/501100011033, the Catalan project 2021 SGR 01559, and the PhD grant AGAUR (2023 FI-3-00223). The Computer Vision Center is part of the CERCA Program/Generalitat de Catalunya.

References

1. Biescas, N., Boned, C., Lladós, J., Biswas, S.: Geocontrastnet: Contrastive key-value edge learning for language-agnostic document understanding. In: International Conference on Document Analysis and Recognition. pp. 294–310. Springer (2024) 3
2. Biswas, S., Riba, P., Lladós, J., Pal, U.: Beyond document object detection: instance-level segmentation of complex layouts. International Journal on Document Analysis and Recognition (IJDAR) **24**(3), 269–281 (2021) 3
3. Carbonell, M., Riba, P., Villegas, M., Fornés, A., Lladós, J.: Named entity recognition and relation extraction with graph neural networks in semi structured documents. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9622–9627. IEEE (2021) 3
4. Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.L., Huang, H., Zhou, M.: Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. NAACL (2021) 7
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: 58th Annual Meeting of the Association for Computational Linguistics (2020) 2, 3, 7
6. Cui, L., Xu, Y., Lv, T., Wei, F.: Document ai: Benchmarks, models and applications. arXiv preprint arXiv:2111.08609 (2021) 1
7. Dai, H.S., Li, X.H., Yin, F., Yan, X., Mei, S., Liu, C.L.: Graphmllm: a graph-based multi-level layout language-independent model for document understanding. In: International Conference on Document Analysis and Recognition. pp. 227–243. Springer (2024) 3
8. Das, A., Biswas, S., Banerjee, A., Lladós, J., Pal, U., Bhattacharya, S.: Harnessing the power of multi-lingual datasets for pre-training: Towards enhancing text spotting performance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 718–728 (2024) 9

9. Davis, B., Morse, B., Cohen, S., Price, B., Tensmeyer, C.: Deep visual template-free form parsing. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 134–141. IEEE (2019) 3
10. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wiginton, C.: Visual fudge: Form understanding via dynamic graph editing. In: International Conference on Document Analysis and Recognition. pp. 416–431. Springer (2021) 1, 3
11. Gemelli, A., Biswas, S., Civitelli, E., Lladós, J., Marinai, S.: Doc2graph: a task agnostic document understanding framework based on graph neural networks. In: European Conference on Computer Vision. pp. 329–344. Springer (2022) 1, 2, 3, 4, 7
12. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017) 1, 2
13. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6. IEEE (2019) 3, 6
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) 1, 3
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013) 2, 7
16. Qasim, S.R., Mahmood, H., Shafait, F.: Rethinking table recognition using graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 142–147. IEEE (2019) 3
17. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019) 2, 3, 7
18. Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., Lladós, J.: Table detection in invoice documents by graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 122–127. IEEE (2019) 3
19. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017) 1
20. Wang, D., Ma, Z., Nourbakhsh, A., Gu, K., Shah, S.: Docgraphlm: Documental graph language model for information extraction. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1944–1948 (2023) 1
21. Wang, J., Jin, L., Ding, K.: Lilt: A simple yet effective language-independent layout transformer for structured document understanding. arXiv preprint arXiv:2202.13669 (2022) 3
22. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. arXiv preprint arXiv:2104.08836 (2021) 3, 6, 7
23. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: Xfund: A benchmark dataset for multilingual visually rich form understanding. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 3214–3224 (2022) 3, 6
24. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? Advances in neural information processing systems **34**, 28877–28888 (2021) 9