Doc2GraphFormer: Bridging Structured Graph Learning with Transformer Attention for Efficient Document Understanding

Souparni Mazumder[§2⋆], Sanket Biswas[§1], Aniket Pal[3], Alloy Das[4], Umapada Pal[5], Josep Lladós[1]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[2] Netaji Subhash Engineering College, Kolkata, India
[3] CVIT Lab, IIIT Hyderabad, India
[4] Habitat Labs, Habitat Lens Pvt. Ltd., India
[5] CVPR Unit, Indian Statistical Institute, Kolkata, India

Abstract. Structured document understanding requires capturing both semantic relationships and spatial structures within visually rich documents. While Graph Neural Networks (GNNs) effectively model structured dependencies, they struggle with long-range dependencies and global context awareness, which are strengths of transformers. In this work, we propose Doc2GraphFormer, a hybrid framework that integrates graph-based structured parsing with multi-head attention mechanisms, evolving from GraphSAGE-style message passing to a more expressive graph-enhanced transformer architecture. Additionally, we introduce a novel modality fusion strategy, inspired by LayoutLMv3, to enhance the integration of textual, visual, and structural embeddings. We evaluate Doc2GraphFormer on FUNSD, XFUND, demonstrating its superior performance in semantic entity recognition and relation extraction. Our findings highlight the potential of combining graph reasoning with self-attention for efficient document understanding. Our code will be made available on: https://github.com/biswassanket/doc2graphformer

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

1 Introduction

Understanding documents is a fundamental challenge in Document AI [5], 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. Graph Neural Networks (GNNs) [11,15,30] have been widely adopted for document structure representation [8,10,33], effectively modeling inter-entity relationships between geometric, textual and visual components. However, most of these approaches rely

⋆ Work done during Internship at Computer Vision Center
§ Equal contribution

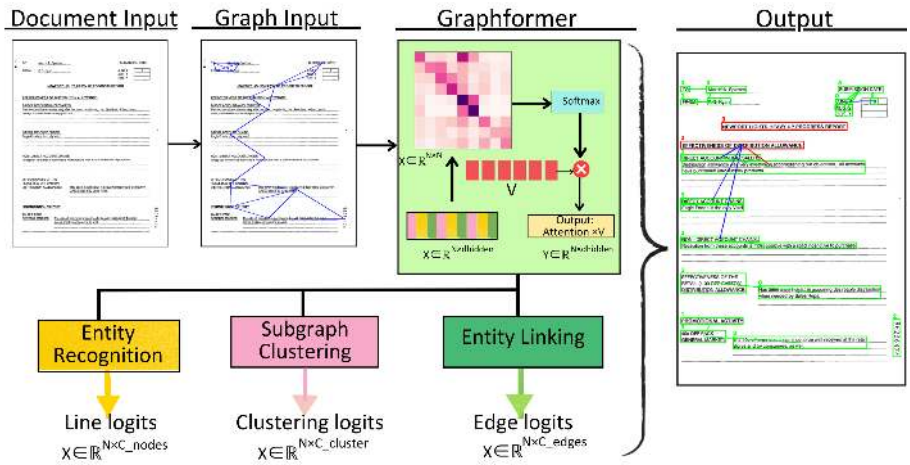

Fig. 1: **Overview of the Doc2GraphFormer Pipeline.** The model converts documents into graph representations, applies self-attention for structured reasoning, and optimizes three tasks: Entity Recognition (classifies entities), Subgraph Clustering (groups related entities), and Entity Linking (predicts key-value relationships). The final output highlights detected entities (green) and entity relationships (blue), showcasing effective structured document understanding.

on GraphSAGE-based message passing mechanisms [11] which struggle with long-range dependencies and global context modeling, limiting their ability to capture complex document structures. On the other hand, transformer-based models [9,13,19,36,37], excel in capturing global relationships via self-attention mechanisms [29] but lack an explicit structured representation of document elements. They process text in a sequential token-based manner, meaning they analyze words in linear order. However, structured documents (e.g., forms, invoices, and reports) contain logically grouped entities that may be spatially distant in the text sequence but closely related in meaning. For example, in an invoice, the "Invoice Total" label may appear at the top of the document, while the corresponding "$5000" value could be at the bottom. A token-based model might struggle to capture this relationship, whereas a graph-based model can directly connect them using edges to reflect their structured relationship.

To overcome these limitations, an ideal solution should integrate the structured reasoning of GNNs with the global context modeling of transformers. We introduce `Doc2GraphFormer`, a hybrid graph-transformer framework that incorporates structured parsing within a multi-head attention architecture [29]. Unlike GraphSAGE-based GNNs[11], which rely on localized message passing, Graph Attention Networks (GATs)[30], which emphasize local neighborhood aggregation, or transformer-based models, which process documents in a sequential token-based manner, Doc2GraphFormer dynamically refines entity relationships using self-attention, capturing both local and global dependencies. This

enables the model to propagate structured information more effectively across the document graph, ensuring robust key-value extraction, relation modeling, and subgraph clustering in structured document understanding. Additionally, we propose a novel multimodal fusion strategy, inspired by LayoutLMv3 [13], to integrate text, visual, and geometric features into a unified graph representation. This work contributes the first structured graph-transformer fusion framework for document parsing and introduces a benchmark study on FUNSD [14,32] and a partial case study on multilingual XFUND [38] benchmark. We introduce three task-specific heads designed for structured document understanding: Semantic Entity Recognition (SER), Subgraph Clustering, and Entity Linking (EL) as shown in Figure 1. Additionally, it leverages grouping labels from the recently released R-FUND benchmark [18], which provide explicit entity grouping annotations. By incorporating these grouping constraints, our approach significantly enhances SER performance, enabling more accurate detection of multi-line key-value pairs and structured entities. The contributions of this work can be summarized as follows: (1) *Graph-Augmented Transformer Learning* → We integrate graph-based reasoning with global self-attention, allowing for adaptive relationship modeling in structured documents. (2) *Multimodal Feature Fusion* → Inspired by LayoutLMv3 [13], we unify textual, visual, and geometric features into a coherent graph representation for enhanced document parsing. (3) *Task-Specific Heads for Structured Parsing* → We introduce SER, Subgraph Clustering, and EL modules, enabling fine-grained key-value extraction and relation modeling.

2 Related Work

Graph Neural Networks (GNNs) were initially introduced for Document Layout Analysis (DLA) tasks [3,1,4], including table detection [24] and table structure recognition [22]. Their adoption was motivated by two primary factors: (i) the ability to leverage the geometric properties of documents for structural analysis, and (ii) the preservation of textual content privacy, particularly in administrative documents, by ensuring that models remain language-independent and structure-driven [24,25,2]. For instance, GNN-based table detection in invoices [24] demonstrated the feasibility of processing documents effectively without reliance on raw text. Building upon this concept, Carbonell et al.[15] for form analysis in the FUNSD benchmark [14], addressing word-level entity grouping, labeling, and linking. Their approach utilized bounding box information and word embeddings as node features while defining edge connections through k-nearest neighbors (KNNs). However, this method's absence of visual features limited its ability to capture layout-based relationships. To overcome this limitation, the FUDGE framework [8] incorporated relationship pairs by employing a lightweight CNN-based detection model [7], followed by a GCN. This integration enhanced text entity classification and key-value pair extraction. Further advancing the field, Doc2Graph [10] introduced an end-to-end, task-agnostic GNN approach for structured document understanding, addressing both node classifi-

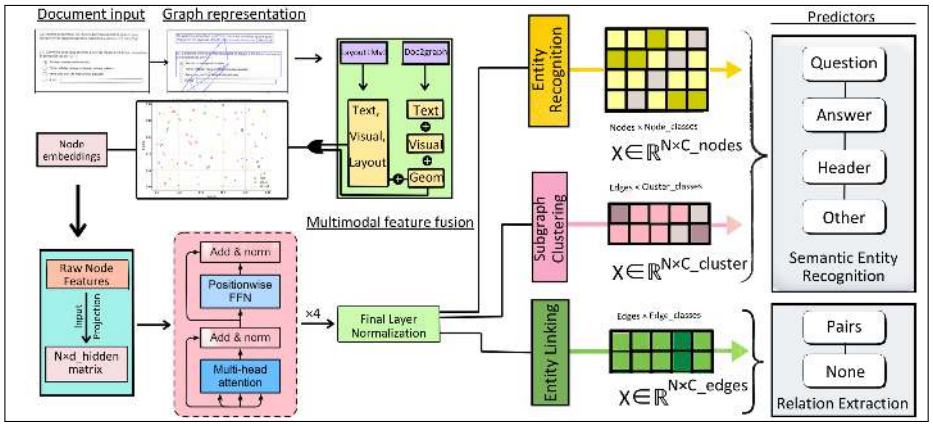

Fig. 2: **Overview of the Doc2GraphFormer Architecture.** Given a structured document, we first construct a graph representation where nodes represent semantic entities and edges capture their relationships. The multimodal feature fusion module integrates textual, visual, and layout embeddings from LayoutLMv3 [13] or Doc2Graph [10]. The resulting node embeddings are then processed using a lightweight transformer encoder with self-attention, for three different task heads as depicted.

cation (SER task) and edge classification (RE task). However, a key limitation of Doc2Graph was its monolingual nature, restricting its applicability to multilingual document processing. Wang et al. [35] presented a method that represents each PDF page as a structured graph, framing the DLA problem as a graph segmentation and classification task. Their Graph-based Layout Analysis Model (GLAM) is a lightweight GNN that competes with state-of-the-art models on challenging DLA datasets while being significantly smaller in size. More recently, models such as StrucTexT [17], UDOP [28], and LiLT [34] have further advanced the field by incorporating structural and cross-lingual pretraining strategies. Le et al. [18] propose a unified pipeline for line extraction, line grouping, and entity linking, which Reduces error accumulation and effectively manages multi-line entities. In this work, we introduce a hybrid lightweight graph-transformer that combines structured reasoning with self-attention. It dynamically refines entity relationships, incorporates multimodal features, and employs task-specific heads for SER, subgraph clustering, and relation extraction, achieving state-of-the-art performance on FUNSD [14] among the lightweight structured reasoning models.

3 Proposed Methodology

In this section, we introduce **Doc2GraphFormer**, a hybrid graph-transformer framework designed for structured document understanding. Unlike conventional Graph Neural Networks (GNNs) that rely on localized message passing, or trans-

formers that process text sequentially, Doc2GraphFormer integrates *graph reasoning* with *global self-attention* to dynamically model inter-entity relationships as shown in Figure 2. The framework consists of three key components: (1) *Graph Construction*, (2) *Graphformer-Based Feature Processing*, and (3) *Task-Specific Heads* for Semantic Entity Recognition (SER), Subgraph Clustering, and Relation Extraction (RE).

3.1 Document Graph Representation

We represent a document as an *undirected graph* $G = (V, E)$, where:

- V is the set of nodes representing document entities (e.g., words, phrases, table cells).
- E is the set of edges encoding relationships between entities.

Each node $v_i \in V$ is associated with a multimodal feature vector:

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

where:

- $\mathbf{t}_i$ represents *textual embeddings* (from SBERT or LayoutLMv3),
- $\mathbf{v}_i$ captures *visual features* (from U-Net or LayoutLMv3),
- $\mathbf{l}_i$ encodes *layout information* (2D bounding box positions from LayoutLMv3),
- $\mathbf{g}_i$ models *geometry-aware features* (which encodes relative positioning with polar coordinates as in Doc2Graph [10]).

3.2 Graph Construction and Attention Masking

In Doc2GraphFormer, we construct a fully connected graph $G = (V, E)$, where each node is initially connected to all other nodes in the document. Instead of relying on predefined edges such as K-Nearest Neighbors (KNN), the model dynamically learns entity relationships by refining connections through the self-attention mechanism. This allows the network to selectively prioritize meaningful relationships while suppressing irrelevant connections.

To incorporate structure-aware attention, we define an *adaptive attention mask*:

$$\mathbf{A}_{ij} = \begin{cases} \text{softmax}(\mathbf{Q}_i \mathbf{K}_j^T) & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{Q}_i$ and $\mathbf{K}_j$ are query and key vectors obtained from Graphformer's self-attention layers. By learning to predict relationships dynamically, our method eliminates the need for manual graph construction heuristics, ensuring that graph connectivity is optimized end-to-end based on semantic, spatial, and structural relevance.

3.3 Graphformer-Based Feature Processing

At the core of Doc2GraphFormer is the **Graphformer encoder**, consisting of *stacked self-attention layers* with position-aware updates.

Each node embedding is refined using:

$$\mathbf{h}_i^{(l+1)} = \text{LayerNorm}\left(\mathbf{h}_i^{(l)} + \text{FFN}\left(\text{MultiHead}(\mathbf{h}_i^{(l)})\right)\right) \tag{3}$$

where:

- **MultiHead** is the multi-head self-attention function,
- **FFN** is a position-wise feedforward network.

This enables *direct long-range dependency modeling*, overcoming the bottleneck of iterative message passing in GNNs.

3.4 Task-Specific Heads

To achieve structured document parsing, Doc2GraphFormer employs three task-specific heads: (1) *Semantic Entity Recognition (SER)*, (2) *Subgraph Clustering*, and (3) *Relation Extraction (RE)*. Each head is optimized independently but shares the same underlying node representations.

Semantic Entity Recognition (SER) Given a node $v_i \in V$, its representation $h_i \in \mathbb{R}^{d_{\text{hidden}}}$ is computed from the Graphformer encoder. The node classification head predicts a label $\hat{y}_i$ from a predefined category set C_{node} (e.g., *Header, Key, Value, Other* in form-based datasets). The logits for classification are computed as:

$$\hat{y}_i = \text{softmax}(W_{\text{SER}} h_i + b_{\text{SER}}) \tag{4}$$

where:

- $W_{\text{SER}} \in \mathbb{R}^{d_{\text{hidden}} \times C_{\text{node}}}$ is a learnable weight matrix.
- $b_{\text{SER}} \in \mathbb{R}^{C_{\text{node}}}$ is the bias vector.

The classification loss is computed using the standard cross-entropy loss:

$$\mathcal{L}_{SER} = -\sum_{i \in V} y_i \log \hat{y}_i \tag{5}$$

where y_i is the ground-truth entity label.

Subgraph Clustering (Grouping Entities) Structured documents contain entities that are logically related but may not be explicitly linked (e.g., multi-line key-value pairs). The subgraph clustering head learns to group related nodes by predicting a binary grouping score for each edge.

For an edge $(i, j) \in E$, we compute:

$$z_{g,ij} = \text{ReLU}\left(W_{\text{group}_1} \begin{bmatrix} h_i \\ h_j \end{bmatrix} + b_{\text{group}_1}\right) \tag{6}$$

where:

- $W_{\text{group}_1} \in \mathbb{R}^{2d_{\text{hidden}} \times d_{\text{hidden}}}$ projects concatenated node embeddings into hidden space.
- $b_{\text{group}_1} \in \mathbb{R}^{d_{\text{hidden}}}$ is the bias vector.

The final grouping logits are computed as:

$$\hat{g}_{ij} = \sigma(W_{\text{group}_2} z_{g,ij} + b_{\text{group}_2}) \tag{7}$$

where:

- $W_{\text{group}_2} \in \mathbb{R}^{d_{\text{hidden}} \times 2}$ maps to a binary grouping decision.
- $b_{\text{group}_2} \in \mathbb{R}^2$ is the bias vector.

The clustering loss is computed using binary cross-entropy:

$$\mathcal{L}_{\text{Cluster}} = -\sum_{(i,j) \in E} g_{ij} \log \hat{g}_{ij} + (1 - g_{ij}) \log(1 - \hat{g}_{ij}) \tag{8}$$

where g_{ij} is the ground-truth grouping label.

Relation Extraction (Entity Linking) To identify explicit relationships between entities, the relation extraction head predicts whether an edge represents a valid link (e.g., key-value pairs in structured forms).

For each edge $(i, j) \in E$, the model computes:

$$z_{r,ij} = \text{ReLU}\left(W_{\text{rel}_1} \begin{bmatrix} h_i \\ h_j \end{bmatrix} + b_{\text{rel}_1}\right) \tag{9}$$

where:

- $W_{\text{rel}_1} \in \mathbb{R}^{2d_{\text{hidden}} \times d_{\text{hidden}}}$ projects concatenated node embeddings into hidden space.
- $b_{\text{rel}_1} \in \mathbb{R}^{d_{\text{hidden}}}$ is the bias vector.

The final relation logits are given by:

$$\hat{r}_{ij} = \sigma(W_{\text{rel}_2} z_{r,ij} + b_{\text{rel}_2}) \tag{10}$$

where:

- $W_{\text{rel}_2} \in \mathbb{R}^{d_{\text{hidden}} \times 2}$ maps to a binary classification decision.
- $b_{\text{rel}_2} \in \mathbb{R}^2$ is the bias vector.

The relation extraction loss is:

$$\mathcal{L}_{\text{RE}} = -\sum_{(i,j) \in E} r_{ij} \log \hat{r}_{ij} + (1 - r_{ij}) \log(1 - \hat{r}_{ij}) \tag{11}$$

where r_{ij} is the ground-truth relation label.

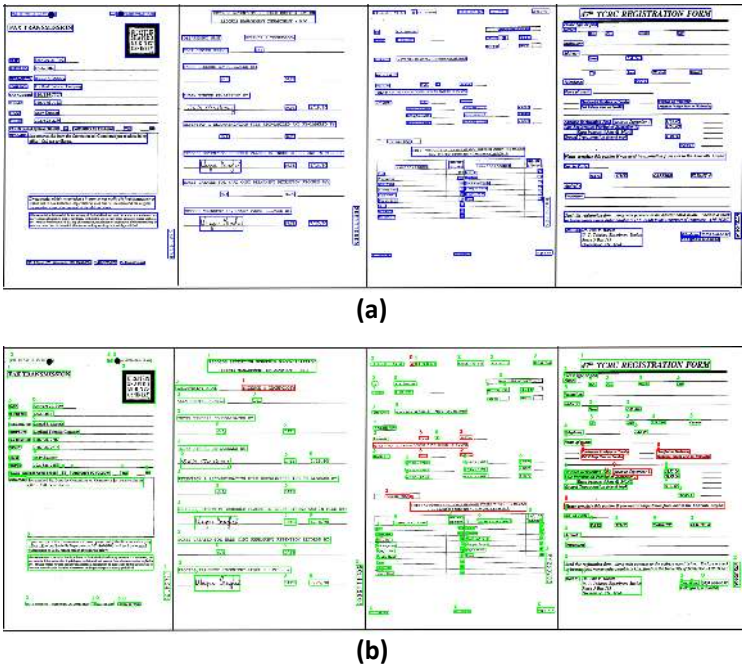

Fig. 3: **Semantic Entity Recognition (SER) Performance Comparison.** (a) Ground truth annotations with labeled entities. (b) Predicted results from Doc2GraphFormer, where green boxes indicate correctly detected entities, and red boxes highlight incorrect predictions. The model effectively captures structured information but struggles with certain misclassified or missing entities, showcasing areas for improvement in handling complex layouts.

3.5 Final Learning Objective

The total training loss is a weighted combination of the three tasks:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{SER} + \lambda_2 \mathcal{L}_{Cluster} + \lambda_3 \mathcal{L}_{RE} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the contributions of each task.

4 Experimental Validation

In this section, we provide a detailed evaluation of Doc2GraphFormer, focusing on its Semantic Entity Recognition (SER) and Relation Extraction (RE) performance. We compare our approach with state-of-the-art models, conduct ablation studies to assess key model components, and present both quantitative and qualitative insights.

Table 1: Comparison of state-of-the-art methods for Semantic Entity Recognition (SER) and Relation Extraction (RE). "GNN" and "Transformer" indicate the usage of Graph Neural Networks and Transformer-based architectures.

Method	**Modalities**	**Graph**	**Transformer**	**SER**	**RE**	**# Params** ($\times 10^6$)
BROS [12]	T + V	✗	✓	0.8121	0.6696	138
LayoutLM [37]	T + V	✗	✓	0.7895	0.4281	343
FUNSD [14]	T + G	✓	✗	0.5700	0.0400	-
FUDGE [8]	V + G	✓	✗	0.6507	0.5241	12
Doc2Graph [10]	T + G + V	✓	✗	0.8225	0.5336	6.2
Voutharoja et al. [**?**]	G	✓	✗	0.8225	0.8540	0.0000081
GeoContrastNet [2]	G + V	✓	✓	0.6476	0.3245	14
Doc2GraphFormer	T + G + V	✓	✓	**0.8439**	**0.5548**	**3.62**
Doc2GraphFormer+GL	T + G + V	✓	✓	**0.8617**	**0.5548**	**3.62**

Table 2: Ablation study on different modality combinations. SER and RE denote Semantic Entity Recognition and Relation Extraction scores.

Text	**Visual**	**Layout**	**Geom**	**SER**	**RE**
✗	✗	✗	✓	0.4077	0.0165
✗	✓	✓	✓	0.6589	0.0801
✓	✗	✓	✓	0.8418	0.5109
✓	✓	✗	✓	0.6991	0.1094
✓	✓	✓	✗	0.8366	0.5138
✓	✓	✓	✓	0.8439	0.5548

4.1 Datasets and Evaluation Metrics

The evaluation is conducted on two widely used structured document understanding benchmarks: `FUNSD` [14] and `XFUND` [38]. FUNSD consists of English-scanned forms annotated for key-value entity linking. XFUND extends FUNSD to a multilingual setting, covering seven additional languages: `Chinese, French, Japanese, German, Italian, Spanish, and Portuguese`. Both datasets provide entity-level annotations and relationship links, making them suitable for evaluating *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. We evaluated Doc2GraphFormer on both Semantic Entity Recognition (SER) and Relation Extraction (RE) tasks using multiple configurations on XFUND. In addition to our baseline model, we conducted experiments with variants that use Sentence-BERT (SBERT) [23] and LayoutLMv3 [13] for feature extraction. In our experiments, the graph trans-

Table 3: Ablation study on the impact of edge weights and attention masks on Semantic Entity Recognition (SER) and Relation Extraction (RE).

Edge weights	Attention mask	SER	RE
✗	✓	0.8418	0.5022
✓	✗	0.8396	0.5046
✓	✓	0.8439	0.5548

Table 4: Ablation study on task contributions. SER and RE denote Semantic Entity Recognition and Relation Extraction scores.

Entity Rec.	Subgraph Clus.	Grouping Labels	Entity Link	SER	RE
✓	✗	✗	✓	0.8426	0.5109
✓	✓	✗	✗	0.8418	0.5022
✓	✓	✗	✓	0.8439	0.5548
✓	✓	✓	✓	0.8617	0.5548

former backbone remains unchanged while the input features vary according to the extraction method.

4.2 State-of-the-Art Competitors

Table 1 presents the F1 scores for SER and RE across multiple languages. Despite having only *3.62 M parameters*, Doc2GraphFormer achieves performance comparable to state-of-the-art transformer-based models, such as LayoutLM [37] (343M parameters) and BROS [12] (138M parameters). BROS and LayoutLM primarily rely on textual and visual features (T+V) while leveraging transformer-based architectures. BROS performs well on SER (0.8121 F1-score) due to its bidirectional text modeling but struggles with RE (0.6696 F1-score) due to the lack of explicit structural reasoning. LayoutLM, despite its multimodal nature, shows lower scores in both SER (0.7895) and RE (0.4281), likely due to the absence of explicit graph-based relational modeling. Graph-Based Models (FUDGE [8], Doc2Graph [10], Voutharoja et al. [31]) introduce explicit structural reasoning, helping to improve RE performance. Doc2Graph [10] stands out among them as it incorporates textual, geometric, and visual information (T+G+V) and achieves strong SER (0.8225) and RE (0.5336) performance. Voutharoja et al., although purely relies on graph-based reasoning (G), achieves the highest RE (0.8540) but is not scalable owing to the fact that it is based on certain heuristic rule-based algorithm. GeoContrastNet [2] is an attempt to combine graph learning with visual U-Net [27] but lags behind in SER (0.6476) and RE (0.3245) due to no language learning modality available. Our approach Doc2GraphFormer integrates text, geometric, and visual cues (T+G+V) within a hybrid Graph-Transformer framework. It outperforms all other methods, achieving state-of-the-

art SER (0.8617) and competitive RE (0.5548) with only **3.62M** parameters, significantly fewer than LayoutLM (343M) and Doc2Graph (6.2M).

4.3 Ablation Studies

Effect of Different Modalities The ablation study on Table 3 highlights the impact of different modality combinations on Semantic Entity Recognition (SER) and Relation Extraction (RE), emphasizing the necessity of multimodal fusion in document understanding. Text alone is insufficient, as indicated by the low SER (0.4077) and RE (0.0165) scores, demonstrating that additional modalities are crucial. Incorporating visual and layout features significantly enhances performance, boosting SER to 0.6589 and RE to 0.0801, showcasing the importance of spatial and appearance-based cues. Adding geometric features further strengthens SER (0.8418), as structured spatial reasoning helps in identifying entities more effectively. While Text + Visual + Geometric slightly reduces SER (0.8366), it improves RE (0.5138), reinforcing the importance of geometric features for relationship extraction. Ultimately, the full multimodal configuration (Text + Visual + Layout + Geometric) achieves the highest scores (0.8439 SER, 0.5548 RE), validating that integrating all modalities leads to the most robust document understanding. This analysis underscores the role of geometric and layout-aware processing, particularly in relation extraction, and highlights the advantages of multimodal learning in Document AI.

Impact of Edge Features and Attention Masks The ablation study investigates the impact of edge weights and attention masks on Semantic Entity Recognition (SER) and Relation Extraction (RE). The results demonstrate that using only attention masks (without edge weights) already achieves a strong SER score of 0.8418 and RE of 0.5022, suggesting that attention-based mechanisms play a key role in capturing entity information. When only edge weights are enabled (without attention masks), there is a slight drop in SER (0.8396) but a small improvement in RE (0.5046), indicating that edge weights contribute to better relational reasoning. The best performance is obtained when both edge weights and attention masks are combined, leading to the highest SER (0.8439) and RE (0.5548). These results highlight that edge weighting helps refine relational modeling, while attention masks enhance entity representation, and their combination leads to more accurate and structured document understanding.

Impact of Different Task Heads The ablation study in Table 4 analyzes the impact of different task contributions—Entity Recognition, Subgraph Clustering, Grouping Labels, and Entity Linking—on Semantic Entity Recognition (SER) and Relation Extraction (RE) performance. The results indicate that Entity Recognition alone with Entity Linking achieves an SER score of 0.8426 and RE of 0.5109, serving as a strong baseline. When Subgraph Clustering is added, the SER remains similar at 0.8418, but RE slightly drops to 0.5022, suggesting that clustering alone does not significantly improve relation extraction. The

Table 5: **Fine-tuning F1 score on XFUND** (fine-tuning on individual language X, testing on X). "SER" denotes semantic entity recognition and "RE" denotes relation extraction.

	Fusion Strategy	ZH	FR
SER	SBERT	**70.02**	**78.95**
	LayoutLMv3	65.39	73.75
RE	SBERT	29.18	27.83
	LayoutLMv3	**34.21**	**33.36**

Table 6: Comparison of different multimodal fusion strategies for SER and RE. The best-performing setting is highlighted

	F_1 (↑)	
Model	**SER**	**RE**
Doc2Graph [10]	0.8210	0.2929
$\text{Doc2Graph}_{\text{SBERT}}$	0.8188	0.3250
LayoutLMv3 [13]	**0.8439**	**0.5548**

introduction of Entity Linking further boosts performance, with SER reaching 0.8439 and RE improving to 0.5548, indicating that structured entity linking plays a crucial role in document understanding. The best results are obtained when all components, including Grouping Labels, are incorporated, achieving the highest SER score of 0.8617 while maintaining a strong RE score of 0.5548. This confirms that Grouping Labels enhance entity recognition, while Entity Linking strengthens relational modeling, leading to more structured and contextually aware document parsing.

Performance on Multiple Languages Table 5 presents a comparison of fine-tuning F1 scores on XFUND for Semantic Entity Recognition (SER) and Relation Extraction (RE) across two languages: Chinese (ZH) and French (FR), using different fusion strategies (SBERT vs. LayoutLMv3). For SER, SBERT outperforms LayoutLMv3 in both languages, achieving 70.02 in ZH and 78.95 in FR, compared to 65.39 and 73.75, respectively. This suggests that SBERT provides more effective semantic embeddings for entity recognition. However, for RE, LayoutLMv3 significantly outperforms SBERT, with 34.21 (ZH) and 33.36 (FR), compared to 29.18 and 27.83 for SBERT. This indicates that LayoutLMv3 better captures relational dependencies between entities, likely due to its multimodal nature, which considers text, layout, and visual information together. The trade-off between SER and RE performance suggests that different fusion strategies may be better suited for different tasks, and combining these techniques could further enhance structured document understanding.

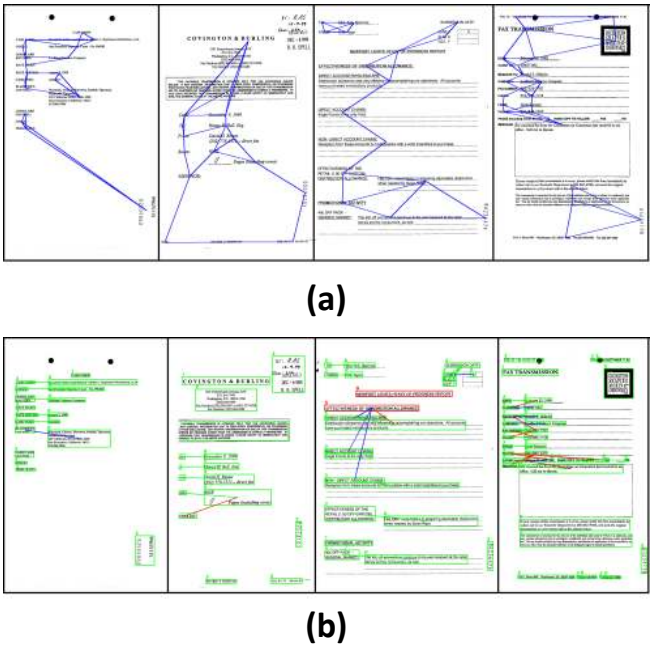

Fig. 4: **SER and RE Performance Comparison.** (a) Ground truth annotations with entity labels and relationships. (b) Predicted results, where green boxes denote correct entities, red boxes highlight entity classifications, and blue lines represent predicted links. The model successfully captures structured dependencies and entity relationships in documents.

Effect of Multimodal Fusion Strategies Table 6 compares different multimodal fusion strategies for Semantic Entity Recognition (SER) and Relation Extraction (RE). LayoutLMv3 achieves the best performance, with an SER score of 0.8439 and an RE score of 0.5548, outperforming both Doc2Graph and Doc2Graph$_{\text{SBERT}}$. This highlights the effectiveness of LayoutLMv3's multimodal feature fusion, which integrates text, layout, and visual cues. Doc2Graph$_{\text{SBERT}}$ improves RE performance (0.3250 vs. 0.2929 in Doc2Graph), indicating that leveraging SBERT embeddings enhances relational understanding between document entities. Doc2Graph performs competitively in SER (0.8210), but lags in RE (0.2929), showing that while its graph-based representation aids entity recognition, it struggles to capture relationships as effectively as transformer-based fusion methods. These results confirm that hybrid transformer-graph models, particularly those integrating strong multimodal fusion mechanisms, lead to better overall document understanding.

4.4 Qualitative Analysis

Figure 3 and Figure 4 provide qualitative examples of SER and RE predictions respectively.
- For *Semantic Entity Recognition (SER)*, Doc2GraphFormer task results are consistent across 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 where closely spaced characters lead to slight misclassification of tokens. Often overlapping or dense content areas can lead to fragmented entity detections, though overall detection remains robust.
- *Relation Extraction (RE)* Analysis as depicted in Figure 4 presents the entity linking (relation extraction) results, where the model predicts semantic relationships between detected entities. We observe that the model correctly identifies key-value pair relationships, even in different writing directions. It can handles nested entities, linking fields correctly within tabular structures. There are minimal false positives, ensuring most connections are meaningful and aligned with document semantics. The edge ambiguity in complex tables, where relationships are less explicit due to layout variations or missing contextual cues. Despite these challenges, Doc2GraphFormer consistently boosts capturing document structure, leveraging graph-based reasoning to improve both entity recognition and relation extraction.

5 Conclusion and Future Work

In this work, we presented Doc2GraphFormer—a multimodal document understanding framework that integrates advanced feature extraction with a graph transformer backbone. By fusing textual, layout, and visual information, our model successfully extracts structured representations from documents, enabling accurate semantic entity recognition and relation extraction. Our experiments on FUNSD and XFUND(fr/zh) demonstrate that leveraging advanced textual embeddings (from Sentence-BERT and LayoutLMv3), along with explicit grouping supervision derived from RFUND-en annotations, leads to significant performance improvements over baseline methods. The ablation studies further confirm that each modality and the grouping signal play a critical role in enhancing the overall system accuracy. For future work, we plan to explore more robust fusion techniques for combining modalities and investigate alternative transformer architectures to further improve efficiency and performance. In addition, we aim to extend our framework to accommodate a wider variety of document types and languages as in [21,6], as well as incorporate instruction-tuned [26] or self-supervised methods [16,20] to better leverage large-scale unlabeled data. Finally, integrating more "layout-aware" grouping and linking signals could further refine the extracted structure and improve downstream tasks in document analysis.

Acknowledgements

The authors acknowledge the financial support of the Department of Research and Universities of the Generalitat of Catalonia to the DocAI Research Group: Group on Document Intelligence (2021 SGR 01559), Grant PID2021-126808OB-I00 (GRAIL) and Grant CNS2022-135947 (DOLORES) funded by MCIN/AEI/10.13039/501100011033, and by ERDF/EU and Ph.D. Scholarship from AGAUR (2023 FI-3-00223). The Computer Vision Center is part of the CERCA Program/Generalitat de Catalunya.

References

1. Banerjee, A., Biswas, S., Lladós, J., Pal, U.: Graphkd: Exploring knowledge distillation towards document object detection with structured graph creation. In: International Conference on Document Analysis and Recognition. pp. 354–373. Springer (2024) 3
2. 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, 9, 10
3. 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
4. Biswas, S., Riba, P., Lladós, J., Pal, U.: Graph-based deep generative modelling for document layout generation. In: International Conference on Document Analysis and Recognition. pp. 525–537. Springer (2021) 3
5. Cui, L., Xu, Y., Lv, T., Wei, F.: Document ai: Benchmarks, models and applications. arXiv preprint arXiv:2111.08609 (2021) 1
6. 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) 14
7. 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
8. 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, 9, 10
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019) 2
10. 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, 3, 4, 5, 9, 10, 12
11. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017) 1, 2

12. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: 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. vol. 36, pp. 10767–10775 (2022) 9, 10
13. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022) 2, 3, 4, 9, 12
14. 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, 4, 9
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) 1, 3
16. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: Dit: Self-supervised pre-training for document image transformer. In: Proceedings of the 30th ACM international conference on multimedia. pp. 3530–3539 (2022) 14
17. Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., Ding, E.: Structext: Structured text understanding with multi-modal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 1912–1920. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475345, https://doi.org/10.1145/3474085.3475345 4
18. Lin, Z., Wang, J., Li, T., Liao, W., Huang, D., Xiong, L., Jin, L.: Peneo: Unifying line extraction, line grouping, and entity linking for end-to-end document pair extraction. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 5171–5180. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3680931, https://doi.org/10.1145/3664647.3680931 3, 4
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019) 2
20. Maity, S., Biswas, S., Manna, S., Banerjee, A., Lladós, J., Bhattacharya, S., Pal, U.: Selfdocseg: A self-supervised vision-based approach towards document segmentation. In: International Conference on Document Analysis and Recognition. pp. 342–360. Springer (2023) 14
21. Mazumder, S., Biswas, S., Das, A., Lladós, J.: Doc2graph-x: A multilingual graph-based framework for form understanding. In: International Workshop on Graph-Based Representations in Pattern Recognition. pp. 257–266. Springer (2025) 14
22. 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
23. 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) 9
24. 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
25. Riba, P., Goldmann, L., Terrades, O.R., Rusticus, D., Fornés, A., Lladós, J.: Table detection in business document images by message passing networks. Pattern Recognition **127**, 108641 (2022) 3

26. Rodriguez, J.A., Jian, X., Panigrahi, S.S., Zhang, T., Feizi, A., Puri, A., Suresh, A.K., Savard, F., Masry, A., Nayak, S., et al.: Bigdocs: An open dataset for training multimodal models on document and code tasks. In: The Thirteenth International Conference on Learning Representations (2025) 14
27. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015) 10
28. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19254–19264 (2023). https://doi.org/10.1109/CVPR52729.2023.01845 4
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 2
30. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017) 1, 2
31. Voutharoja, B.P., Qu, L., Shiri, F.: Language independent neuro-symbolic semantic parsing for form understanding. arXiv preprint arXiv:2305.04460 (2023) 10
32. Vu, H.M., Nguyen, D.T.N.: Revising funsd dataset for key-value detection in document images. arXiv preprint arXiv:2010.05322 (2020) 3
33. 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
34. 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) 4
35. Wang, J., Krumdick, M., Tong, B., Halim, H., Sokolov, M., Barda, V., Vendryes, D., Tanner, C.: A graphical approach to document layout analysis. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023. pp. 53–69. Springer Nature Switzerland, Cham (2023) 4
36. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020) 2
37. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: 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. pp. 1192–1200 (2020) 2, 9, 10
38. 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, 9