KANNER .

Zhe Li, K. L. Eddie Law
Faculty of Applied Sciences, *Macao Polytechnic University*
Macao, China
{zhe.li, eddielaw}@mpu.edu.mo

***Abstract*—Grid-based methods have become the dominant paradigm for Named Entity Recognition (NER) by explicitly enumerating all possible spans. However, these methods face a fundamental trade-off between global context aggregation and local boundary precision. Conventional convolutional encoders effectively capture long-range dependencies but inadvertently act as low-pass filters, blurring the sharp boundaries required for accurate span detection. Furthermore, traditional Multi-Layer Perceptrons (MLPs) often struggle to disentangle highly coupled semantic features in nested structures without excessive parameter growth. To resolve these issues, we propose KANNER, a parameter-efficient architecture that integrates Kolmogorov-Arnold Networks (KANs) into the NER framework. Specifically, we design a parallel enhancement mechanism: (1) Multi-scale Adaptive Gradient Sharpening (MAGS), which employs learnable difference operators to recover high-frequency boundary signals from the smoothed context; and (2) Masked Spectral KAN Attention (MaskedSpeKAN), which leverages the learnable activation functions of KANs to capture non-linear channel dependencies with fewer parameters than standard MLPs. Experiments on CADEC, CoNLL2003, and GENIA benchmarks demonstrate that KANNER achieves state-of-the-art performance. Notably, it surpasses strong baselines by +1.01 F1-score on CADEC while maintaining competitive inference speeds, offering a superior balance between accuracy and efficiency.**

***Index Terms*—Depthwise separable convolutions, grid-based methods, Kolmogorov-Arnold Networks, Named Entity Recognition, span-based methods.**

I. Introduction

Named Entity Recognition (NER) [1] is a core task in information extraction that aims to identify and classify named entities such as persons, organizations, locations, and other predefined categories within unstructured text. As a fundamental component of natural language processing (NLP), NER has wide-ranging applications, including information retrieval, question answering, text summarization, and machine translation. For instance, in the sentence "*The endothelin receptor antagonist is a medication that reduces protein in urine for patients with kidney disease*," a well-trained NER model is expected to identify and categorize relevant medical terms such as drug names and diseases with high accuracy [2].

Depending on the structural relationships between entities, NER tasks are commonly categorized into flat and nested entity recognition [3]. While flat entities appear as non-overlapping spans, nested entities exhibit hierarchical structures where one entity is contained within another. Recognizing these structures is critical in specialized domains like biomedical text mining and legal document analysis. Recently, grid-based methods (often utilizing biaffine mechanisms) have emerged as the dominant paradigm for Nested NER [4], [5]. These approaches transform the sentence into an $N \times N$ matrix to exhaustively enumerate all possible spans, theoretically resolving the overlapping issue. Despite their success, existing grid-based paradigms face two fundamental challenges that hinder further performance gains:

1) *Smoothing dilemma of local context*: To aggregate contextual information across the grid, prior models frequently employ Convolutional Neural Networks (CNNs). While smoothing strengthens contextual semantics, it acts as a low-pass filter, inevitably expanding the receptive field and sacrificing local signal sharpness. This phenomenon blurs the decision boundary between valid entities and background spans, resulting in "boundary insensitivity" [6].
2) *Linear bottleneck of feature space*: Conventional attention mechanisms and classifiers typically rely on Multi-Layer Perceptrons (MLP) with fixed activation functions. However, the semantic features of nested entities are often highly entangled and non-convex. Simple linear mappings struggle to disentangle noise from core semantic signals in high-dimensional spaces, creating a bottleneck for precision [7].

In this paper, we propose KANNER, a novel architecture that integrates Kolmogorov-Arnold Networks (KAN) [8] into the NER framework, to tackle these challenges. Unlike traditional feed-forward neural networks that use fixed activation functions on nodes, KANs place learnable nonlinear activation functions (B-splines) on edges. Based on this mathematical property, we construct a parallel feature enhancement mechanisms:

- **Spatial Sharpening via MAGS**: To resolve the "smoothing dilemma," we introduce the Multi-scale Adaptive Gradient Sharpening (MAGS) module. Instead of standard smoothing kernels, MAGS utilizes difference operators initialized as high-pass filters. This allows the model to recover sharp boundary signals from the smoothed context, effectively distinguishing entity boundaries from the background.
- **Channel Purification via MaskedSpeKAN**: To break the "linear bottleneck," we propose the Masked Spectral KAN Attention (MaskedSpeKAN) module. By replacing linear MLPs with non-linear KAN layers, this module

This work was supported in part by Macao Polytechnic University (MPU) Grant RP/FCA-05/2025 with an internal document ID fca.8057.8277.5

captures complex dependencies between feature channels. Furthermore, it incorporates a masking mechanism to explicitly handle padding noise in variable-length sequences, ensuring robust feature purification.

We empirically validate KANNER on widely used nested NER benchmarks. Our model achieves state-of-the-art performance. It demonstrates superior effectiveness in handling complex nested structures. Our key contributions are summarized as follows:

1) We identify the "*smoothing dilemma*" and "*linear bottleneck*" in existing grid-based NER models and propose KANNER, the first architecture to integrate Kolmogorov-Arnold Networks into Nested NER to address these issues.
2) We design a parallel enhancement layer comprising MAGS for spatial boundary sharpening and MaskedSpeKAN for non-linear channel denoising, significantly improving boundary precision and feature discriminability.

II. Related Works

A. Span Extraction

Recent advancements show that transforming NER into a span classification task effectively recognizes nested entities. The boundary-enhanced neural span classification model introduced in [9] added a boundary detection task to identify entity boundaries, improving nested named entity recognition through a multi-task learning framework. Span representations were further enhanced using span-level graphs [10], which connected spans and entities with n-gram features to better recognize low-frequency entities. In [11], various aspects of span-based NER were examined, including representations, learning strategies, and decoding algorithms to avoid overlaps, with an efficient algorithm proposed for finding non-overlapping span sets to maximize global scores. The SpanKL model [12] was introduced as a simple yet effective approach for continuous learning in NER, preserving memory through knowledge distillation and using multi-label predictions to avoid conflicts.

B. Sequence Labeling and Generative Approaches

Although grid-based and span-based methods currently dominate the Nested NER landscape, sequence labeling remains a vital research direction due to its high inference efficiency. Traditional CRF-based models with standard tagging schemes (e.g., BIO, BIOES) struggle to represent nested structures. However, recent works in 2023 and 2024 have revitalized this paradigm. Some approaches propose novel tagging schemes, such as layered tagging or anchor-region labeling, to map hierarchical entities into linear sequences without collision [13]. More notably, with the rise of Large Language Models (LLMs), the definition of sequence labeling has expanded to include generative paradigms. Recent studies leverage instruction tuning to guide LLMs (e.g., LLaMA, T5) to output linearized entity sequences directly, treating NER as a sequence-to-sequence generation task [14]. While these generative methods show promise in few-shot scenarios, they often suffer from inference latency and hallucination issues compared to dedicated lightweight architectures like KANNER.

C. Hypergraph-based Methods

To explicitly model the complex interactions between overlapping spans, hypergraph neural networks (HGNNs) have been introduced to the NER domain. Unlike simple graph-based methods that only capture pairwise dependencies (word-word or span-span), hypergraphs utilize hyperedges to connect multiple nodes, thereby capturing high-order semantic correlations [15]. For instance, recent models construct hypergraphs where nodes represent text spans and hyperedges represent potential entity clusters or boundary interactions [16]. By performing message passing on these hyperstructures, models can effectively resolve the ambiguity of nested boundaries. However, the construction of dynamic hypergraphs is often computationally expensive and sensitive to noise in the graph structure. In contrast, our KANNER avoids the heavy overhead of graph construction by employing the MAGS module to sharpen boundaries directly within the feature grid.

D. Kolmogorov-Arnold Networks

Unlike traditional Multi-Layer Perceptrons (MLPs), Kolmogorov-Arnold Networks (KANs) [8] belong to a new type of neural network architecture that utilizes learnable activation functions on the edges instead of fixed activation functions on the nodes. This allows KANs to replace linear weights with univariate functions, often represented as B-spline interpolations, enhancing flexibility and interpretability.

Different from MLPs which applying fixed non-linear activation functions, KANs dynamically adjust the transformations based on the data. Each weight in KAN is a learnable activation function which improves adaptability. Furthermore, compared to MLPs, KANs usually achieve high accuracy with fewer parameters. As a result, the architecture is less prone to overfitting, and more suitable for tasks in scientific machine learning.

III. Methodology

In this section, we present the architecture of KANNER, as illustrated in Figure 1. The model adopts a grid-based paradigm and consists of four primary components:

1) a grid construction layer based on BERT and the biaffine mechanism;
2) a context aggregation layer utilizing Depthwise Separable CNN;
3) a parallel feature enhancement layer comprising MAGS and MaskedSpeKAN; and
4) a dynamic fusion and decision layer based on LSGF and KAN.

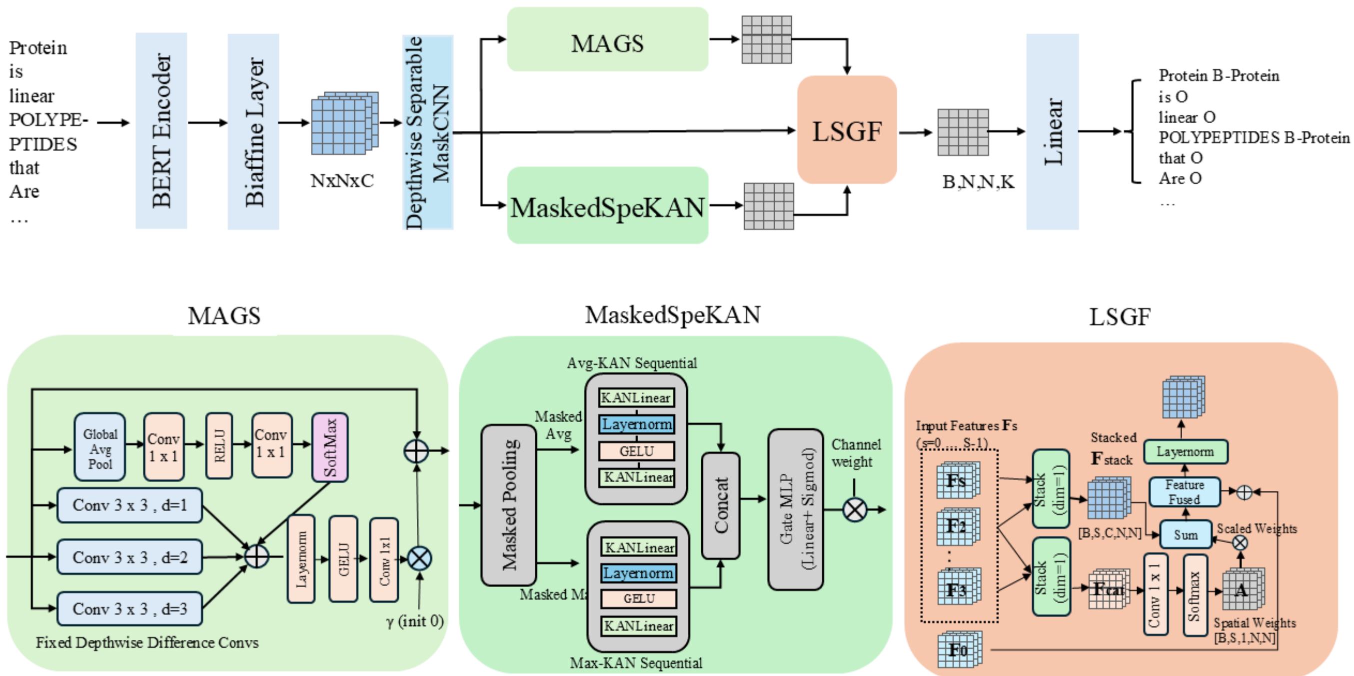

Fig. 1. KANNER model architecture.

A. Problem Formulation

Given an input sentence $X = \{x_1, x_2, \ldots, x_N\}$ consisting of N tokens, the task of Nested NER is to extract a set of entity triples $\mathcal{E} = \{(i, j, c) \mid 1 \leq i \leq j \leq N, c \in \mathcal{Y}\}$, where i and j denote the start and end indices, and $c \in \mathcal{Y}$ is the category. We formulate this as a pixel-wise classification problem on an $N \times N$ grid, where the feature vector at grid point (i, j) represents the semantic attributes of span $x_{i:j}$.

B. Encoder and Grid Construction

a) Contextual Encoding: We employ a pre-trained language model to extract hidden states $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\} \in \mathbb{R}^{N \times d_{\text{model}}}$. A max-pooling strategy is applied to aggregate sub-token embeddings into word-level representations.

b) Biaffine Mechanism: To construct the grid, we use a Biaffine mechanism to model word-pair interactions. First, two MLPs project $\mathbf{H}$ into head and tail representations:

$$\mathbf{h}_i^{(s)} = \text{MLP}_{\text{head}}(\mathbf{h}_i), \tag{1}$$

$$\mathbf{h}_j^{(e)} = \text{MLP}_{\text{tail}}(\mathbf{h}_j). \tag{2}$$

Then, we compute biaffine scores to generate the initial grid $\mathbf{V}_{\text{raw}} \in \mathbb{R}^{N \times N \times d_{\text{out}}}$:

$$\mathbf{V}_{\text{raw}}^{(i,j)} = \mathbf{h}_i^{(s)} \mathbf{U} (\mathbf{h}_j^{(e)})^\intercal + \mathbf{W}(\mathbf{h}_i^{(s)} \oplus \mathbf{h}_j^{(e)}) + \mathbf{b}, \tag{3}$$

where $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{b}$ are trainable parameters, and $\oplus$ denotes concatenation.

C. Depthwise Separable CNN

To efficiently smooth features, we introduce a Depthwise Separable MaskCNN. Let $\mathbf{V}_{\text{base}}$ denote the smoothed base features:

$$\mathbf{V}_{\text{base}} = \text{PWConv}(\text{DWConv}(\mathbf{V}_{\text{raw}} \odot \mathbf{M}_{\text{pad}})), \tag{4}$$

where $\mathbf{M}_{\text{pad}}$ is a mask matrix to shield padding regions. $\mathbf{V}_{\text{base}}$ serves as the backbone for subsequent enhancement branches.

D. Parallel Feature Enhancement

1) Multi-scale Adaptive Gradient Sharpening (MAGS): To mitigate boundary blurring, MAGS sharpens semantic edges in the spatial dimension. Let $\mathcal{D} = \{1, 2, 3\}$ be the set of dilation rates. For each $d \in \mathcal{D}$, we compute the sharpened feature $\mathbf{F}_d$. The final boundary-enhanced feature $\mathbf{V}_{\text{mags}}$ is:

$$\mathbf{V}_{\text{mags}} = \mathbf{V}_{\text{base}} + \gamma \cdot \sum_{d \in \mathcal{D}} \alpha_d \cdot \text{Conv}_d(\mathbf{V}_{\text{base}}), \tag{5}$$

where γ is a learnable scaling parameter initialized to 0, and α_d are scale weights computed via a lightweight attention mechanism.

2) Masked Spectral KAN Attention (MaskedSpeKAN): [Image of Kolmogorov-Arnold Network structure vs Multi-Layer Perceptron] Targeting redundancy in channels, MaskedSpeKAN leverages KAN's non-linear fitting capability. The masked global mean z_c for the c-th channel is:

$$z_c = \frac{\sum_{i,j} \mathbf{V}_{\text{base}}^{(i,j,c)} \cdot \mathbf{M}_{\text{valid}}^{(i,j)}}{\sum_{i,j} \mathbf{M}_{\text{valid}}^{(i,j)} + \epsilon}. \tag{6}$$

We then employ KAN layers to model inter-channel dependencies, generating weights $\omega \in \mathbb{R}^{d_{\text{out}}}$:

$$\omega = \sigma(\text{KAN}_{\text{max}}(\text{MaxPool}(\mathbf{V}_{\text{base}})) + \text{KAN}_{\text{avg}}(\mathbf{z})), \quad (7)$$

where KAN layers consist of learnable B-spline functions. The enhanced feature is $\mathbf{V}_{\text{spekan}} = \mathbf{V}_{\text{base}} \odot \omega + \mathbf{V}_{\text{base}}$.

E. Dynamic Fusion and Span Classification

a) Lightweight Spatial Gated Fusion (LSGF): We construct a feature set $\mathcal{F} = \{\mathbf{V}_{\text{base}}, \mathbf{V}_{\text{raw}}, \mathbf{V}_{\text{mags}}, \mathbf{V}_{\text{spekan}}\}$. The LSGF module allocates weights via:

$$\mathbf{A} = \text{Softmax}(\text{Conv}_{1\times1}([\mathcal{F}])) \in \mathbb{R}^{N\times N\times|\mathcal{F}|}. \quad (8)$$

The final fused feature $\mathbf{V}_{\text{fused}}$ is:

$$\mathbf{V}_{\text{fused}} = \sum_{k=1}^{|\mathcal{F}|} \mathbf{A}_k \odot \mathcal{F}_k + \mathbf{V}_{\text{base}}. \quad (9)$$

b) Span Classification: Finally, the probability distribution over entity labels is computed as:

$$P(y_{i,j}|X) = \text{Softmax}(\mathbf{W}_{\text{cls}}\mathbf{V}_{\text{fused}}^{(i,j)} + \mathbf{b}_{\text{cls}}), \quad (10)$$

where $\mathbf{W}_{\text{cls}}$ and $\mathbf{b}_{\text{cls}}$ are learnable weights and bias.

IV. Experiments

A. The Datasets

To validate the effectiveness of our model, we selected the CADEC [18],CoNLL-2003 datasets [19] and Genia [20]. The information about these datasets is shown in the Table I. These datasets are selected due to the following three key criteria: 1) freely available, 2) open-source, include natural language explanations for labels, and 3) are widely recognized. Genia is a collection of biomedical literature, containing 1,999 Medline abstracts and features nested named entities. CADEC (Corpus of Adverse Drug Event Conversations) is a specialized biomedical dataset focused on adverse drug events (ADEs), derived from online patient forums. It contains entity annotations for drug names, diseases, symptoms, and side effects, making it particularly valuable for clinical named entity recognition (NER) tasks. The dataset presents significant challenges due to informal language, spelling variations, and the diverse terminology of medical concepts, making it a crucial resource for biomedical NLP and pharmacovigilance research. In contrast, CoNLL03 (Conference on Natural Language Learning 2003) is one of the most widely used benchmark datasets for general NER, comprising newswire text in English, German, Spanish, and Dutch. It includes entity labels for persons (PER), organizations (ORG), locations (LOC), and miscellaneous entities (MISC). Originally introduced as part of the CoNLL 2003 shared task, CoNLL03 remains a standard benchmark for evaluating NER models, playing a pivotal role in the advancement of deep learning and pre-trained language models in NLP.Genia is a collection of biomedical literature, containing 1,999 Medline abstracts and features nested named entities. Genia primarily has five labels, which are Protein, Cell type, Cell line, DNA, and RNA.

B. Implementation Details

Our model is implemented based on the PyTorch framework. To ensure optimal semantic representation across different domains, we select specific pre-trained language models as the backbone encoder:

- General Domain: For the CoNLL03 dataset, we utilize the standard BERT-large model.
- Biomedical Domain: For the CADEC and GENIA datasets, we employ BioBERT-Large. Pre-trained on large-scale biomedical corpora (PubMed and PMC), BioBERT-Large effectively captures complex medical terminology and dependencies that are challenging for general-purpose models.

We use the AdamW optimizer for training. The detailed hyperparameter settings are provided in the Appendix.

TABLE I
THE STATISTICAL DATA OF THE DATASET

Data Split	Entity Count		
	CADEC	***CoNLL03***	***GENIA***
Train	8401	28,073	15,023
Dev	1096	3,508	1,669
Test	1148	3,508	1,854

C. Performance on Nested Datasets

1) Results on CADEC: Table II presents the performance comparison on the CADEC dataset. Our model achieves a remarkable F1-score of **75.05**, establishing a new state-of-the-art result. Specifically, it outperforms the previous best model TOE by a significant margin of **1.01**. In terms of detailed metrics, KANNER achieves a recall of **74.22** and a precision of **75.90**. Compared to TOE, while our precision is slightly lower, our balanced improvement in recall demonstrates that KANNER effectively retrieves more valid nested entities without drastically sacrificing accuracy. This improvement is largely attributed to the MAGS module, which successfully captures the subtle boundary signals of nested spans that are often missed by traditional smoothing methods.

2) Results on GENIA: Table V outlines the results on the GENIA corpus, another widely used benchmark for nested NER. KANNER achieves a state-of-the-art F1-score of **81.84**, outperforming the diffusion-based approach (DiffusionNER) by **0.31** points. Notably, while Triaffine-NER achieves the highest recall (82.06), our model dominates in precision with a score of **82.76**. This trade-off suggests that in highly dense biomedical texts, the MaskedSpeKAN module is particularly effective at filtering out false positives by distinguishing fine-grained entity types (e.g., distinguishing *DNA* from *RNA* based on subtle context cues). The consistent improvements across both CADEC and GENIA, powered by **BioBERT-Large**, validate the robustness of KANNER in handling complex hierarchical structures in the biomedical domain.

D. Performance on the Flat Dataset

Table IV compares the performance of our model with leading baselines on the standard flat NER dataset, CoNLL2003. Our model achieves a top F1-score of **94.75**, surpassing all compared methods. Notably, KANNER shows an improvement of **+1.58** F1-score over the W2NER baseline (93.17) and outperforms other recent approaches like DiFiNet (+1.03 F1-score). Although CoNLL2003 contains only flat entities, the superior performance of our model confirms that the proposed feature enhancement mechanisms are not limited to nested structures. The MaskedSpeKAN module effectively purifies feature channels, allowing the model to focus on core semantic signals even in general-domain news texts, thereby boosting generalization ability.

E. Results and Discussion

The experimental results across both biomedical (nested) and general (flat) domains highlight two key strengths of KANNER:

1) Robustness Across Domains: Unlike specialized models that excel only in specific scenarios, KANNER demonstrates consistent gains in diverse environments. On the CADEC dataset, which features complex nested structures and informal clinical language, the **+1.01** F1-score gain over TOE validates our hypothesis that KAN-based non-linearity is crucial for modeling entangled semantic features. Simultaneously, the performance on CoNLL2003 indicates that our architecture preserves the linear separability of simple flat entities while enhancing boundary detection.

2) Addressing the Smoothing Dilemma: The quantitative improvements in boundary-sensitive metrics (Precision and Recall) support the effectiveness of our parallel enhancement strategy. Standard CNN-based models often suffer from "boundary insensitivity" due to over-smoothing. By incorporating the Multi-scale Adaptive Gradient Sharpening (MAGS) module, our model recovers high-frequency spatial details. This is qualitatively reflected in the model's ability to distinguish closely overlapping spans in the CADEC dataset, offering a robust solution to the "Smoothing Dilemma" identified in the Introduction.

In summary, KANNER successfully balances the need for rich contextual semantics with the requirement for precise boundary delineation, offering a promising solution for complex information extraction tasks.

TABLE II
PERFORMANCE COMPARISON OF NER MODELS ON THE CADEC DATASET.

Model	Performance Metrics		
	F1-Score	*Recall (R)*	*Precision (P)*
[21]	71.5	72.5	70.5
Caio [22]	72.9	/	/
TOE [23]	74.04	70.66	**77.77**
W2NER [4]	73.21	72.35	74.09
[24]	72.95	70.11	76.03
Ours	**75.05**(+1.01)	**74.22**	75.9

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT SETTINGS ON THE CADEC DATASET.

Settings	Performance Metric
	F1-Score
KBC	**75.05**
w/o MAGS	74.45
w/o MaskedSpeKAN	74.72

TABLE IV
PERFORMANCE COMPARISON OF NER MODELS ON THE CONLL2003 DATASET. † MEANS THAT OUR RE-IMPLEMENTATION VIA THEIR CODE.

Model	Performance Metrics		
	F1-Score	*Recall*	*Precision*
W2NER†	93.17	94.22	92.17
DiffusionNER [25]	92.78	92.56	92.99
PromptNER [26]	92.41	92.33	92.48
BINDER [27]	93.33	93.57	93.08
DiFiNet [28]†	93.72	93.60	93.84
Ours	**94.75** (+1.03)	**94.36**	**95.14**

F. Ablation Study and Model Efficiency

To investigate the contribution of each component to the overall performance, we conducted an ablation study on the CADEC dataset, as shown in Table III.

1) Impact of Components: The results clearly validate the necessity of our proposed modules. When the **MAGS** module is removed (w/o MAGS), the F1-score drops significantly by **0.70** (from 75.05 to 74.35). This decline highlights the critical role of gradient sharpening in resolving the "Smoothing Dilemma," as the model loses the ability to recover sharp boundary signals from the smoothed context. Similarly, removing the **MaskedSpeKAN** module (w/o MaskedSpeKAN) leads to a performance decrease of **0.33** F1-score (74.72). This demonstrates that the non-linear channel purification provided by KAN is effective in filtering out redundant features and noise, although the spatial sharpening (MAGS) appears to have a slightly more dominant impact on nested entity recognition.

2) Computational Efficiency: Beyond predictive performance, KANNER is designed for computational efficiency. Grid-based methods typically suffer from high computational costs due to the $N \times N$ feature map. However, our incorporation of Depthwise Separable Convolutions (DWCNN) drastically reduces the number of parameters and floating-point operations (FLOPs) compared to standard dense CNNs used

TABLE V
PERFORMANCE COMPARISON OF NER MODELS ON THE GENIA DATASET. † MEANS THAT OUR RE-IMPLEMENTATION VIA THEIR CODE.

Model	Performance Metrics		
	F1-Score	*Recall (R)*	*Precision (P)*
Biaffine-NER	80.5	79.3	81.8
CNN-NER	80.33	79.17	81.52
DiffusionNER	81.53	80.97	82.1
Triaffine-NER	81.23	**82.06**	80.42
Ours	**81.84**(+0.31)	80.95	**82.76**

in prior works. By decoupling spatial and channel correlations, DWCNN maintains a large receptive field with a fraction of the computational burden. Furthermore, while KAN introduces learnable activation functions, recent studies [**?**] suggest that KANs possess better parameter efficiency than MLPs, often achieving comparable fitting power with fewer parameters. This combination allows KANNER to achieve state-of-the-art accuracy without imposing prohibitive hardware requirements, making it suitable for deployment in resource-constrained scenarios.

V. Conclusion

In this paper, we propose a KANNER model, which is a grid-based approach for carrying out Named Entity Recognition (NER) operations. The architecture of KANNER is based on Kolmogorov-Arnold Networks (KANs), which place learnable non-linear activation functions (B-splines) on edges. KANNER is consisted of a pair of feasture enhancement mechanisms: In associating the fundamental trade-off between global context aggregation and local boundary precision with grid-based methods, we discovered the *smoothing dilemma* and *linear bottleneck* issues. We further introduce two module designs, Multi-scale Adaptive Gradient Sharpening (MAGS) and Masked Spectral KAN Attention (MaskedSpeKAN), in KANNER to subside the discovered problematic issues. Through thorough experiments on CADEC, CoNLL2003, and GENIA datasets, KANNER handles complex nested nested data structures and outperforms other state-of-the-art designs. Apart from maintaining competitive inference speeds, KANNER offers a superior balance between accuracy and efficiency.

Appendix

The optimal hyperparameter settings for each dataset in our experiments are shown in Table VI below.

TABLE VI
Hyperparameter settings across datasets

Hyperparameter Settings	**Dataset**		
	Genia	***CADEC***	***Conll03***
Batchsize	8	8	8
Epoch	8	15	10
Learning rate	7e-6	5e-6	5e-6
Biaffine size	120	120	120
CNN channel dim	120	120	200
Dropout rate	0.1	0.1	0.1

References

[1] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History," in *Proc. 16th Int. Conf. Comput. Linguistics (COLING 1996)*, vol. 1, 1996, pp. 466–471.

[2] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022.

[3] Y. Wang, H. Tong, Z. Zhu, and Y. Li, "Nested Named Entity Recognition: A Survey," *ACM Trans. Knowl. Discov. Data*, vol. 16, no. 6, Art. 108, Dec. 2022.

[4] J. Li et al., "Unified Named Entity Recognition as Word-Word Relation Classification," *arXiv preprint arXiv:2112.10070*, 2021.

[5] H. Yan et al., "A Unified Generative Framework for Various NER Subtasks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 5808–5822.

[6] Y. Zhang, Y. Xie, and E. Xing, "Span-based NER with BERT," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL 2019)*, 2019, pp. 4433–4442.

[7] X. Han and J. Zhao, "A survey of nested named entity recognition: Datasets, evaluation methods, and models," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 4–23, 2020.

[8] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljacic, T. Y. Hou, M. Tegmark, "KAN: Kolmogorov-Arnold Networks," *Proc. ICLR*, 2025.

[9] C. Tan, W. Qiu, M. Chen, R. Wang, and F. Huang, "Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020.

[10] J. Wan, D. Ru, W. Zhang, and Y. Yu, "Nested Named Entity Recognition with Span-level Graphs," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL 2022)*, vol. 1, 2022, pp. 892–903.

[11] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Named Entity Recognition as Structured Span Prediction," in *Proc. Workshop Unimodal Multimodal Induction Linguistic Struct. (UM-IoS)*, 2022, pp. 1–10.

[12] Y. Zhang and Q. Chen, "A neural span-based continual named entity recognition model," in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, pp. 1–9.

[13] D. Vilares and C. Gómez-Rodríguez, "Nested Named Entity Recognition as Single-Pass Sequence Labeling," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, FL, USA, Nov. 2024, pp. 9115–9128.

[14] X. Wang et al., "InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction," *arXiv preprint arXiv:2304.08085*, 2023.

[15] Y. Yan, B. Cai, and S. Song, "Nested Named Entity Recognition as Building Local Hypergraphs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, Jun. 2023, pp. 13878–13886.

[16] Z. Zhang, B. Wang, Y. Wang, and U. Luo, "Hypergraph based Understanding for Document Semantic Entity Recognition," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Bangkok, Thailand, Aug. 2024, pp. 2950–2960.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[18] M. Jeong, K. H. Cho, S. Kim, et al., "CADEC: A Large-Scale Clinical Corpus for Named Entity Recognition and Normalization," in *Proc. Workshop Biomed. Nat. Lang. Process.*, 2012, pp. 110–118.

[19] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *Proc. 7th Conf. Nat. Lang. Learn. (CoNLL 2003)*, 2003, pp. 142–147.

[20] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. Suppl 1, pp. i180–i182, Jul. 2003.

[21] J. Yu, B. Bohnet, and M. Poesio, "Named Entity Recognition as Dependency Parsing," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL 2020)*, 2020, pp. 6470–6476.

[22] C. Corro, "A fast and sound tagging method for discontinuous named-entity recognition," *arXiv preprint arXiv:2409.16243*, 2024.

[23] J. Liu et al., "TOE: A Grid-Tagging Discontinuous NER Model Enhanced by Embedding Tag/Word Relations and More Fine-Grained Tags," *arXiv preprint arXiv:2211.00684*, 2022.

[24] T.-C. Chen and W.-Y. Lin, "On Fusing ChatGPT and Ensemble Learning in Discontinuous Named Entity Recognition in Health Corpora," *arXiv preprint arXiv:2412.16976*, 2024.

[25] C. Li, L. Li, X. Huang, et al., "Diffusion-based Named Entity Recognition," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2022)*, 2022, pp. 198–206.

[26] J. Gu, Y. Yang, D. Chen, et al., "PromptNER: A Prompt-based Approach for Named Entity Recognition," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2022)*, 2022, pp. 208–219.

[27] W. Li, R. Zhang, and J. Zhao, "BINDER: A Robust Named Entity Recognition Framework Based on BERT and Knowledge Graph Embeddings," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP 2021)*, 2021, pp. 122–131.

[28] S. Yu, X. Ren, and J. Han, "DiFiNet: A Deep Fusion Network for Named Entity Recognition," in *Proc. NeurIPS 2021*, 2021, pp. 3291–3301.