An Adaptive Multi-Threshold Loss and a General Framework for Collaborating Losses in Document-Level Relation Extraction

Huangming Xu, Fu Zhang*, Jingwei Cheng
School of Computer Science and Engineering, Northeastern University, China
xuhuangming@foxmail.com, zhangfu@mail.neu.edu.cn

Abstract

The goal of document-level relation extraction (DocRE) is to identify relations for a given entity pair within a document. As a multi-label classification task, the most commonly employed method involves introducing an adaptive threshold. Specifically, for an entity pair, if the scores of predicted relations exceed the threshold, the relations exist. However, we observe two phenomena that significantly weaken the model's performance in DocRE: (1) as the label space (the number of relations) expands, the model's performance gradually declines; (2) the model tends to prioritize predicting high-frequency relations in the long-tail problem. To address these challenges, we propose an innovative **A**daptive **M**ulti-**T**hreshold **L**oss (AMTL), which for the first time proposes to partition the label space into different sub-label spaces (thus reducing its overall size) and learn an adaptive threshold for each sub-label space. This approach allows for more precise tuning of the model's sensitivity to diverse relations, mitigating the performance degradation associated with label space expansion and the long-tail problem. Moreover, our adaptive multi-threshold method can be considered as a general framework that seamlessly integrates different losses in different sub-label spaces, facilitating the concurrent application of multiple losses. Experimental results demonstrate that AMTL significantly enhances the performance of existing DocRE models across four datasets, achieving state-of-the-art results. The experiments on the concurrent application of multiple losses with our framework show stable performance and outperform single-loss methods. Code is available at https://github.com/xhm-code/AMTL.

1 Introduction

Document-level relation extraction (DocRE) (Yao et al., 2019) aims to identify one or more relations for an entity pair in documents. Unlike sentence-level relation extraction tasks, DocRE needs to handle longer texts, cross-sentence relations, and more complex contextual dependencies with diverse semantic structures. The increasing demand for accurate relation identification in tasks like question answering (Cao et al., 2022), knowledge graph construction (Ye et al., 2022), and event extraction (Liu et al., 2024) has made DocRE a key area of research in information extraction.

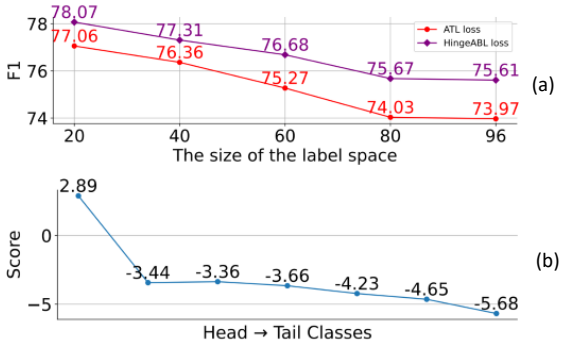

Figure 1: (a) Performance of ATL- and HingeABL-based models as the label space expands. (b) Relation scores for head (high-frequency) and tail (low-frequency) relation classes.

DocRE, as a *multi-label* classification task, traditionally employs binary cross-entropy loss (BCE) to learn a global threshold. Specifically, for each entity pair, if the relation scores exceed the global threshold, the model determines that corresponding relations exist. However, the global threshold frequently fails to adapt to the characteristics of all entity pairs, resulting in suboptimal performance. To address this limitation, Zhou et al. (2021) propose an *adaptive threshold loss* (ATL), which introduces an adaptive threshold for each entity pair so that the scores of positive classes are significantly higher than the threshold, while the scores of negative classes are much lower than the threshold[1].

* Corresponding author.

[1] R is a set of predefined relations in DocRE tasks, where

Findings of the Association for Computational Linguistics: ACL 2025, pages 20996–21007

Although ATL effectively alleviates the problem of the global threshold, it is still insufficient when dealing with the *long-tail* problem[2]. To further address the long-tail problem, Wei and Li (2022) propose an adaptive multi-label loss (AML), while Wang et al. (2023) propose an adaptive hinge balance loss (HingeABL). Both methods are inspired by the hinge loss (Hearst et al., 1998) and enhance the model's ability to distinguish between tail classes and difficult-to-classify relations (where the relation scores are close to the threshold) by widening the gap between positive and negative classes. Similarly, Tan et al. (2022a) propose an adaptive focal loss (AFL) based on the focal loss, which is designed to pay more attention to the tail classes to cope with the performance degradation caused by the long-tail problem.

The above losses use only one adaptive threshold for each entity pair and fail to fully consider the diversity of relations. Specifically, for an entity pair, its relation scores are compared with only one threshold. In the DocRE task, the types of relations are often diverse, with each relation having different semantics. For example, the DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b) datasets include 96 different predefined relations. However, we observe the following two phenomena in the DocRE task: **Firstly**, as shown in **Fig. 1**(a), the performance of models based on the ATL and the state-of-the-art HingeABL decreases significantly as the label space[3] expands. **Secondly**, in the long-tail problem, the models tend to predict the head classes, as shown in **Fig. 1**(b), where the head classes' scores are generally higher than those of the tail classes, which gives the models an excessive preference for the head classes.

To address the performance degradation caused by the expansion of label space and the challenges posed by long-tail problem, we propose a novel multi-label classification loss, the **A**daptive **M**ulti-**T**hreshold **L**oss (AMTL). Specifically, AMTL innovatively proposes to partition the label space into different segments based on the frequency of relation occurrences, and introduces an adaptive threshold for each segment. Our approach effectively reduces the label space size and enables different thresholds for head and tail classes, allowing for precise tuning of the model's sensitivity to diverse relations. Further analysis indicates that the adaptive multi-threshold method serves as a general framework that seamlessly integrates ATL-based losses by introducing them into segments, facilitating the concurrent application of multiple losses. Moreover, our experiments demonstrate that AMTL exhibits effective generalization capabilities; specifically, it can be trained on incompletely labeled datasets while maintaining good prediction performance on fully labeled datasets.

Our contributions are as follows:

- A novel loss, AMTL, is introduced, which for the first time proposes to partition the label space into multiple segments and learn an adaptive threshold for each segment. This effectively mitigates the performance degradation associated with the expansion of label space and the long-tail problem.
- The adaptive multi-threshold method can be considered as a general framework that seamlessly integrates ATL-based losses by introducing them into segments, facilitating the concurrent application of multiple losses.
- The AMTL is thoroughly evaluated on four DocRE datasets, revealing consistent performance enhancements and effective generalization capabilities across various backbone models and achieving state-of-the-art (SOTA) results compared to baseline methods.

2 Related Work

Existing DocRE methods can be broadly divided into the following categories:

(1) **Improvements in Representation Capability.** By designing new model structures or enhancing existing ones, semantic relations in documents are captured more accurately, leading to improved classification performance. For example, GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021), DocuNet (Zhang et al., 2021), KD-DocRE (Tan et al., 2022a), DREEAM (Ma et al., 2023), AA (Lu et al., 2023), SRF (Zhang et al., 2024), REwNCRL (Xu et al., 2024), and TTM-RE (Gao et al., 2024).

(2) **Optimization of Loss.** By designing new losses, the model's performance in DocRE can be

positive classes $\mathcal{P}_T \subseteq R$ represent the relations that exist for an entity pair, while *negative classes* $\mathcal{N}_T \subseteq R$ represent the relations that do not exist, where $R = \mathcal{P}_T \cup \mathcal{N}_T$.

[2]In DocRE datasets, specific relations within the predefined relation set R appear with higher frequency (commonly referred to as *head classes*), while others occur less frequently (referred to as *tail classes*), resulting in the long-tail problem, also known as the class imbalance problem.

[3]*Label space* refers to the number of relations in DocRE.

significantly improved, particularly when addressing complex scenarios such as long-tail problem and multi-label classification. Notable examples include ATL (Zhou et al., 2021) and its extensions, such as Balanced-Softmax (Zhang et al., 2021), AML (Wei and Li, 2022), AFL (Tan et al., 2022a), SSR-PU (Wang et al., 2022), NCRL (Zhou and Lee, 2022), PEMSCL (Guo et al., 2023), and HingeABL (Wang et al., 2023).

(3) **Plugin-based Approach.** Such methods exhibit strong generalization capabilities and can be integrated as modular components into various model architectures, further enhancing model performance. For example, LogicRE (Ru et al., 2021), MILR (Fan et al., 2022), BCBR (Liu et al., 2023), P^{3}M (Wang et al., 2024), and JMRL (Qi et al., 2024). Among these, LogicRE, MILR, and JMRL, as a logical reasoning module, can explicitly capture the long-range dependencies between entities.

3 Methodology

We first define the task of DocRE, followed by an introduction to the most commonly used loss in this task, adaptive threshold loss (ATL). Finally, we present our proposed improvement based on ATL, the adaptive multi-threshold loss (AMTL).

3.1 Problem Formulation

Given a document D and an entity pair $T = (e_s, e_o)$, where e_s is the subject and e_o is the object, the DocRE task is to predict the subset of relations for T from $R \cup \{\text{NA}\}$. Here, R denotes the set of predefined relations, such that $R = \mathcal{P}_T \cup \mathcal{N}_T$, with NA indicating the absence of any relation. The positive classes $\mathcal{P}_T \subseteq R$ are the relations that exist between e_s and e_o; if no relation exists, $\mathcal{P}_T$ is empty. Conversely, the negative classes $\mathcal{N}_T \subseteq R$ are relations that do not exist between e_s and e_o; if there is no relation, $\mathcal{N}_T = R$.

3.2 Adaptive Threshold Loss

The adaptive threshold loss (ATL) (Zhou et al., 2021) is a widely used multi-label classification loss in DocRE. In ATL, as shown in **Eq. (1)**, the set R of predefined relations is divided into two subsets: positive classes $\mathcal{P}_T$ and negative classes $\mathcal{N}_T$. Additionally, ATL introduces a threshold class TH. During training, the loss aims to make the scores of positive classes $\mathcal{P}_T$ significantly higher than the scores of the TH class, and the scores of negative classes $\mathcal{N}_T$ significantly lower than the scores of the TH class. In testing, if the relation scores exceed the TH class, the relations are considered to exist; otherwise, they are assumed not to exist.

$$\mathcal{L}_1 = -\sum_{r \in \mathcal{P}_T} \log \left(\frac{\exp(\text{logit}_r)}{\sum_{r' \in \mathcal{P}_T \cup \{\text{TH}\}} \exp(\text{logit}_{r'})} \right)$$
$$\mathcal{L}_2 = -\log \left(\frac{\exp(\text{logit}_{\text{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\text{TH}\}} \exp(\text{logit}_{r'})} \right) \quad (1)$$
$$\mathcal{L}_{ATL} = \mathcal{L}_1 + \mathcal{L}_2$$

3.3 Adaptive Multi-Threshold Loss

The ATL-based losses employ an adaptive threshold, where relation scores for an entity pair exceeding this threshold are classified as positive classes $\mathcal{P}_T$, while scores below it are deemed negative classes $\mathcal{N}_T$ (Wei and Li, 2022; Tan et al., 2022a; Wang et al., 2023). However, as we have detailed in Section 1, an adaptive threshold fails to fully consider the diversity of relations and cannot adequately address the challenge of model performance degradation caused by label space expansion and the long-tail problem.

To overcome these limitations, we introduce an adaptive multi-threshold loss (AMTL), as shown in **Fig. 2**. Specifically, we rank the label occurrence frequencies in the train set in descending order to differentiate between head and tail classes, partitioning the label space into multiple sub-label space segments. For each sub-label space segment, we set an adaptive threshold to reduce the overall size of the label space. This method allows us to apply different thresholds for head classes and tail classes, effectively alleviating the long-tail problem and label space expansion problem.

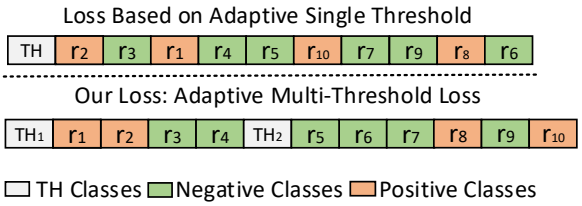

Figure 2: Comparison of ATL-based losses and AMTL.

Moreover, to reduce variability among thresholds and ensure model consistency and stability across TH classes, we apply a weighted average, as shown in **Eq. (2)**. Here, i denotes the i-th sub-label space segment, n represents the number of

sub-label space segments, and λ is the coefficient for the weighted average.

$$\text{logit}^i_{\text{TH}} = \frac{\text{logit}^i_{\text{TH}} + \sum_{j=1, j\neq i}^{n} \text{logit}^j_{\text{TH}}}{\lambda} \quad (2)$$

Using **Eq. (2)**, we compute the loss for the i-th sub-label space segment, which includes contributions from both positive and negative classes, as shown in **Eq. (3)**.

$$\mathcal{L}_3{}^i = -\sum_{r\in\mathcal{P}_T^i} \log\left(\frac{\exp(\text{logit}_r)}{\sum_{r'\in\mathcal{P}_T^i}\exp(\text{logit}_{r'}) + \exp(\text{logit}^i_{\text{TH}})}\right)$$
$$\mathcal{L}_4{}^i = -\log\left(\frac{\exp(\text{logit}^i_{\text{TH}})}{\sum_{r'\in\mathcal{N}_T^i}\exp(\text{logit}_{r'}) + \exp(\text{logit}^i_{\text{TH}})}\right) \quad (3)$$

Finally, we obtain our Adaptive Multi-Threshold Loss (AMTL), as shown in **Eq. (4)**:

$$\mathcal{L}_{AMTL} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}_3^i + \mathcal{L}_4^i\right) \quad (4)$$

3.4 Adaptive Multi-Threshold Framework

Our proposed adaptive multi-threshold method can be viewed as a general framework, which seamlessly integrates ATL-based losses by introducing the same or different losses in different segments as shown in **Fig. 3**, thus facilitating the joint application of multiple losses.

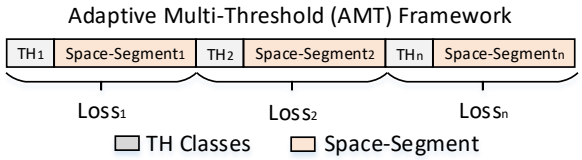

Figure 3: Our adaptive multi-threshold (AMT) general framework for collaborating losses in DocRE.

As shown in **Eq. (5)**, $\mathcal{L}_i$ represents the i-th loss function, which consists of the i-th TH class and the i-th sub-label space segment, n represents the number of sub-label space segments. Furthermore, the update of the i-th TH class can also be performed using **Eq. (2)**.

$$\mathcal{L}_{AMT} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i \quad (5)$$

Dataset	Split	#Docs.	#Rels.	#Triples.
DocRED	train	3,053	96	38,180
	dev	1000	96	12,323
	test	1000	96	-
DWIE	train	602	65	14,403
	dev	98	65	2,624
	test	99	65	2,495
Re-DocRED	train	3,053	96	85,932
	dev	500	96	17,284
	test	500	96	17,448
DocGNRE	train (GPT)	3,053	96	96,505
	train (mGPT)	3,053	96	103,561
	test	500	96	19,526

Table 1: Statistics of datasets.

4 Experimental Settings

4.1 Datasets and Evaluation

We conduct experiments on four datasets: DocRED (Yao et al., 2019), DWIE (Zaporojets et al., 2021), Re-DocRED (Tan et al., 2022b), and DocGNRE (Li et al., 2023), with details provided in **Table 1** and **Appendix B**.

Following Yao et al. (2019), we employ F1 and Ign-F1 as evaluation metrics. The Ign-F1 score is computed by excluding relational facts that are shared between the train and dev/test sets from the F1 calculation.

4.2 Baselines

To conduct a comprehensive comparison and analysis of the superiority of AMTL, we compare it with the **three categories of work** discussed in Section 2. **First**, we evaluate the performance of AMTL on several competitive models, including GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021), DocuNet (Zhang et al., 2021), KD-DocRE (Tan et al., 2022a), DREEAM (Ma et al., 2023), and TTM-RE (Gao et al., 2024). Notably, these baseline models employ different losses: ATLOP and DREEAM use ATL loss, KD-DocRE adopts AFL loss, DocuNet applies Balanced-Softmax loss, and TTM-RE utilizes the SSR-PU loss. We assess the performance of replacing these losses with AMTL loss. **Moreover**, we compare AMTL with existing DocRE losses. **Additionally**, we compare AMTL with plugin-based methods. We also perform a detailed experimental analysis of our AMT framework.

4.3 Implementation Details

We use $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019) as encoders and utilize code from public repositories of various

baseline models for our experiments. To ensure the stability of the results, we conduct experiments with five different random seeds (5, 42, 65, 66, 233) and take the average as the final result. All experiments are conducted on NVIDIA 3090 GPU.

Moreover, we set the number of sub-label space segments to 4 (n = 4) for the DocRED, Re-DocRED, and DocGNRE datasets, and 3 (n = 3) for the DWIE dataset. Based on our experience, the suggested value for the parameter λ is the number of sub-label space segments minus 0.5 for the first three datasets, while for the DWIE dataset, it is recommended to take it as the number of sub-label space segments.

5 Main Results and Analysis

We conduct experiments to answer the following research questions about our main contributions, Adaptive Multi-Threshold Loss (AMTL) and Adaptive Multi-Threshold (AMT) framework:

- **Q1:** How effective is our AMTL loss when applied to different models? (Section 5.1)
- **Q2:** How does the performance of our AMTL loss compare to other losses? (Section 5.2)
- **Q3:** How effective is our adaptive multi-threshold framework AMT compared to existing adaptive single-threshold methods? (Section 5.3)

5.1 Different DocRE Models with AMTL

To evaluate the effectiveness of AMTL loss applied to different models, we replace the losses of various models with AMTL.

As shown in **Table 3**, AMTL consistently improves performance across multiple competitive models on several datasets. Specifically, using RoBERTa$_{\text{large}}$ as the encoder, the F1 score on the Re-DocRED test set improves by an average of 1.85, while Ign-F1 increases by 1.87. Similarly, when DocGNRE (mGPT) is used as the train set, the F1 score on the DocGNRE test set rises by 2.72 on average, with a corresponding gain of 2.64 in Ign-F1.

Moreover, we note that the improvement on the TTM-RE (Gao et al., 2024) model is particularly significant. On the Re-DocRED dev set, F1 increases to 81.72 and Ign-F1 to 80.86, with improvements of 3.59 and 2.81, respectively; on the test set, F1 increases to 82.02 and Ign-F1 to 81.19, with gains of 2.07 and 2.99, respectively, achieving SOTA performance on both dev and test sets.

Loss Combination	Dev		Test	
	F1	Ign-F1	F1	Ign-F1
ATL+AFL+AML+SAT	75.19	73.95	74.75	73.49
ATL+AFL+NCRL+SAT	75.06	73.94	74.59	73.47
ATL+HingeABL+PEMSCL+AFL	75.52	74.29	75.18	73.98
ATL+ATL+AML+AFL	75.43	74.23	75.10	73.94

Table 2: Results of the collaboration of multiple losses with our AMT framework, using ATLOP as the representation module and BERT$_{\text{base}}$ for encoding.

5.2 Different Loss Methods

To evaluate the performance of our loss AMTL against other losses for DocRE, the results presented in **Table 4** demonstrate that AMTL achieves better performance by outperforming all comparative losses in terms of both F1 and Ign-F1 scores.

Specifically, on the Re-DocRED test set, our proposed AMTL achieves an F1 score of 75.63, surpassing the previous best result obtained with HingeABL loss by 0.48. Similarly, the Ign-F1 score reaches 74.44, representing a 0.60 improvement over HingeABL loss. On the DocGNRE test set, AMTL attains an F1 score of 71.34, exceeding HingeABL loss by 0.36, while the Ign-F1 score improves by 0.44 to 70.34. These results indicate that AMTL exhibits superior performance in the DocRE task compared to other losses.

5.3 Multiple Thresholds *vs.* Single Threshold

To demonstrate the effectiveness of our adaptive multi-threshold (AMT) framework compared to existing single-threshold methods, we conducted two sets of experiments. The *first set* uses the same loss in each sub-label space segment within the AMT framework, as shown in **Table 5**, while the *second set* applies different losses in each sub-label space segment, as presented in **Table 2**.

Table 5 shows that the AMT framework consistently improves F1 and Ign-F1 metrics, particularly in methods based on ATL, AML, AFL, and HingeABL$_{SAT}$. For instance, on the Re-DocRED test set, AMT increased ATL's F1 by 2.34 and Ign-F1 by 1.98, while AFL's F1 improved by 1.53 on the dev set. Additionally, **Table 2** demonstrates that combining different losses maintains stable performance and outperforms single-threshold methods, further validating the AMT framework's robustness. However, we did not find a clear pattern for the optimal loss combination, which we plan to explore further.

Model	Dev				Test			
	F1	F1 with AMTL	Ign-F1	Ign-F1 with AMTL	F1	F1 with AMTL	Ign-F1	Ign-F1 with AMTL
Re-DocRED with $\text{BERT}_{\text{base}}$								
ATLOP (Zhou et al., 2021)	73.97 †	**75.80** (+1.83)	73.04 †	**74.59** (+1.55)	73.29	**75.63** (+2.34)	72.46	**74.44** (+1.98)
DocuNet (Zhang et al., 2021)	74.62 △	**76.22** (+1.60)	73.60 △	**74.91** (+1.31)	74.48 △	**76.01** (+1.53)	73.53 △	**74.73** (+1.20)
KD-DocRE (Tan et al., 2022a)	74.66 △	**75.92** (+1.26)	73.68 △	**74.78** (+1.10)	74.55 △	**75.63** (+1.08)	73.64 △	**74.53** (+0.89)
DREEAM (Ma et al., 2023)	74.13 †	**76.10** (+1.97)	73.68 †	**75.34** (+1.66)	73.75 †	**75.71** (+1.96)	73.33 †	**74.98** (+1.65)
TTM-RE (Gao et al., 2024)	75.51 †	**78.97** (+3.46)	74.31 †	**78.03** (+3.72)	75.71 †	**78.72** (+3.01)	74.55 †	**77.81** (+3.26)
Re-DocRED with $\text{RoBERTa}_{\text{large}}$								
ATLOP (Zhou et al., 2021)	77.63 *	**80.09** (+2.46)	76.88 *	**79.15** (+2.27)	77.73 *	**79.97** (+2.24)	76.94 *	**79.04** (+2.10)
DocuNet (Zhang et al., 2021)	78.16 *	**79.62** (+1.46)	77.53 *	**78.54** (+1.01)	77.92 *	**79.71** (+1.79)	77.27 *	**78.66** (+1.39)
KD-DocRE (Tan et al., 2022a)	78.65 *	**79.73** (+1.08)	77.92 *	**78.73** (+0.81)	78.35 *	**79.45** (+1.10)	77.63 *	**78.46** (+0.83)
DREEAM (Ma et al., 2023)	77.60 †	**79.80** (+2.20)	77.20 †	**79.19** (+1.99)	77.94 ◇	**79.98** (+2.04)	77.34 ◇	**79.40** (+2.06)
TTM-RE (Gao et al., 2024)	78.13 ◇	**81.72** (+3.59)	78.05 ◇	**80.86** (+2.81)	79.95 ◇	**82.02** (+2.07)	78.20 ◇	**81.19** (+2.99)
DWIE with $\text{RoBERTa}_{\text{large}}$								
ATLOP (Zhou et al., 2021)	76.65	**77.19** (+0.54)	72.47	**73.13** (+0.66)	81.39	**81.85** (+0.46)	76.83	**77.44** (+0.81)
DocuNet (Zhang et al., 2021) †	76.46	**77.04** (+0.58)	72.69	**72.98** (+0.29)	81.32	**81.55** (+0.23)	77.20	76.91 (-0.29)
KD-DocRE (Tan et al., 2022a) †	76.55	**77.00** (+0.45)	72.01	**73.52** (+1.51)	80.92	**81.05** (+0.13)	75.67	**77.22** (+1.55)
DocGNRE (GPT) with $\text{BERT}_{\text{base}}$								
ATLOP (Zhou et al., 2021)	73.89 †	**75.98** (+2.09)	73.07 †	**74.85** (+1.78)	68.74 †	**71.22** (+2.48)	68.06 †	**70.24** (+2.18)
DREEAM (Ma et al., 2023)	74.23 †	**76.07** (+1.84)	73.76 †	**75.30** (+1.54)	68.24 ‡	**71.28** (+3.04)	68.89 †	**70.63** (+1.74)
TTM-RE (Gao et al., 2024)	75.44 †	**78.93** (+3.49)	74.33 †	**77.96** (+3.63)	71.14 †	**74.02** (+2.88)	70.19 †	**73.19** (+3.00)
DocGNRE (GPT) with $\text{RoBERTa}_{\text{large}}$								
ATLOP (Zhou et al., 2021)	77.61 †	**80.11** (+2.50)	76.96 †	**79.18** (+2.22)	72.90 †	**75.47** (+2.57)	72.36 †	**74.67** (+2.31)
DREEAM (Ma et al., 2023)	77.75 †	**79.74** (+1.99)	77.28 †	**79.14** (+1.86)	72.90 ‡	**75.47** (+2.57)	72.97 †	**74.97** (+2.00)
TTM-RE (Gao et al., 2024)	78.16 †	**81.59** (+3.43)	77.30 †	**80.74** (+3.44)	73.72 †	**77.05** (+3.33)	73.01 †	**76.35** (+3.34)
DocGNRE (mGPT) with $\text{BERT}_{\text{base}}$								
ATLOP (Zhou et al., 2021)	73.80 †	**76.01** (+2.21)	73.00 †	**74.85** (+1.85)	68.81 †	**71.04** (+2.23)	68.16 †	**70.02** (+1.86)
DREEAM (Ma et al., 2023)	74.30 †	**75.92** (+1.62)	73.84 †	**75.15** (+1.31)	68.00 ‡	**71.44** (+3.44)	68.83 †	**70.81** (+1.98)
TTM-RE (Gao et al., 2024)	75.54 †	**78.97** (+3.43)	74.33 †	**78.05** (+3.72)	71.59 †	**74.14** (+2.55)	70.54 †	**73.35** (+2.81)
DocGNRE (mGPT) with $\text{RoBERTa}_{\text{large}}$								
ATLOP (Zhou et al., 2021)	77.70 †	**80.34** (+2.64)	77.02 †	**79.37** (+2.35)	72.99 †	**75.40** (+2.41)	72.44 †	**74.56** (+2.12)
DREEAM (Ma et al., 2023)	77.72 †	**79.89** (+2.17)	77.34 †	**79.27** (+1.93)	73.29 ‡	**75.71** (+2.42)	72.80 †	**75.21** (+2.41)
TTM-RE (Gao et al., 2024)	78.15 †	**81.64** (+3.49)	77.29 †	**80.83** (+3.54)	73.84 †	**77.17** (+3.33)	73.12 †	**76.50** (+3.38)

Table 3: Performance of different DocRE models with AMTL loss. We replace the losses of various models with AMTL. Results with † are our reproduction, ‡ from Qi et al. (2024), * from Lu et al. (2023), △ from Xu et al. (2024), and ◇ from the original paper. For DocGNRE, lacking a dev set, we evaluate using Re-DocRED dev set.

Loss Function	Re-DocRED		DocGNRE	
	F1	Ign-F1	F1	Ign-F1
ATL (Zhou et al., 2021)	73.29 *	72.46 *	68.74 †	68.06 †
Balanced-Softmax (Zhang et al., 2021)	73.68 *	72.85 *	68.84 †	68.13 †
AML (Wei and Li, 2022)	72.60 *	71.78 *	67.86 †	67.11 †
AFL (Tan et al., 2022a)	74.15 *	73.20 *	69.45 †	68.69 †
NCRL (Zhou and Lee, 2022)	73.87 †	72.79 †	69.20 †	68.27 †
SSR-PU (Wang et al., 2022)	73.00 †	71.53 †	69.54 †	68.29 †
PEMSCL (Guo et al., 2023)	73.98 †	73.06 †	69.46 †	68.70 †
HingeABL_{SAT} (Wang et al., 2023)	73.46 *	72.61 *	69.15 †	68.41 †
$\text{HingeABL}_{MeanSAT}$ (Wang et al., 2023)	74.68 *	72.90 *	70.83 †	69.25 †
HingeABL (Wang et al., 2023)	75.15 *	73.84 *	70.98 †	69.90 †
AMTL (Our Loss)	**75.63** (0.48↑)	**74.44** (0.60↑)	**71.34** (0.36↑)	**70.34** (0.44↑)

Table 4: Results of different losses on the Re-DocRED test set (trained on the Re-DocRED train set) and the DocGNRE test set (trained on the DocGNRE (GPT) train set). † indicates our reproduction, and * from Wang et al. (2023). All results use ATLOP (Zhou et al., 2021) as the representation module and employ $\text{BERT}_{\text{base}}$ for encoding.

Loss Function	Dev				Test			
	F1	F1 with AMT	Ign-F1	Ign-F1 with AMT	F1	F1 with AMT	Ign-F1	Ign-F1 with AMT
ATL (Zhou et al., 2021)	73.93†	**75.80** (+1.87)	73.04†	**74.59** (+1.55)	73.29*	**75.63** (+2.34)	72.46*	**74.44** (+1.98)
Balanced-Softmax (Zhang et al., 2021)	74.00†	**74.25** (+0.25)	73.14†	**73.30** (+0.16)	73.68*	**73.89** (+0.21)	72.85*	**72.96** (+0.11)
AML (Wei and Li, 2022)	73.04†	**74.49** (+1.45)	72.17†	**73.38** (+1.21)	72.60*	**73.91** (+1.31)	71.78*	**72.79** (+1.01)
AFL (Tan et al., 2022a)	74.36†	**75.89** (+1.53)	73.36†	**74.63** (+1.27)	74.15*	**75.54** (+1.39)	73.20*	**74.29** (+1.09)
SSR-PU (Wang et al., 2022)	73.57†	**74.06** (+0.49)	72.09†	**72.46** (+0.37)	73.00†	**73.80** (+0.80)	71.53†	**72.20** (+0.67)
PEMSCL (Guo et al., 2023)	74.50†	**75.38** (+0.88)	73.55†	**74.11** (+0.56)	73.98†	**74.93** (+0.95)	73.06†	**73.67** (+0.61)
HingeABL_{SAT} (Wang et al., 2023)	74.06†	**75.47** (+1.41)	73.19†	**74.29** (+1.10)	73.46*	**75.39** (+1.93)	72.61*	**74.25** (+1.64)
HingeABL (Wang et al., 2023)	75.61†	**75.72** (+0.11)	74.39†	**74.54** (+0.15)	75.15*	**75.49** (+0.34)	73.84*	**74.35** (+0.51)

Table 5: Performance comparison of our adaptive multi-threshold framework AMT and existing single threshold methods based on different losses on the Re-DocRED dataset. "**with AMT**" means using our adaptive multi-threshold framework AMT in Fig. 3, where *each sub-label space segment in the framework uses the same loss as the single threshold method*, and $n = 4$ (n represents the number of segments in the framework). All the results use ATLOP (Zhou et al., 2021) as the representation module and employ $\text{BERT}_{\text{base}}$ for encoding. Results marked with † are from our reproduction, * from Wang et al. (2023).

6 Further Analysis

To further investigate our method's performance, we answer the following research questions:

- **Q4:** Does the AMT framework alleviate the long-tail problem? (Section 6.1)
- **Q5:** How effective is our AMTL loss in mitigating performance degradation due to label space expansion? (Section 6.2)
- **Q6:** How does our AMTL loss perform on weakly supervised generalization? (Section 6.3)
- **Q7:** How does our AMTL loss compare to different plugin-based methods regarding performance when integrated into the model? (Section 6.4)
- **Q8:** What is the training cost of our AMTL loss? (Section 6.5)
- **Q9:** How do our AMTL loss and HingeABL loss perform on different models? (Section 6.6)
- **Q10:** Is a larger segment number n always better for datasets with more labels? (Section 6.7)
- **Q11:** What is the impact of hyperparameters (segment number n, weighted average λ, label orderings) on the performance of our AMTL loss? (Section 6.8)

6.1 Analyzing the Long-Tail Problem

In order to analyze the impact of AMT framework on long-tail problem, we first rank all predefined relations in descending order based on their frequency in the Re-DocRED train set. Subsequently, these relations are grouped into four categories: Head-10 (the top 10 relations), Mid-76 (relations ranked 11th to 86th), Tail-20 (the bottom 20 relations), and Tail-10 (the last 10 relations).

As shown in **Table 6**, the AMT framework improves F1 scores across four bands and *effectively alleviate the long-tail problem*. In the Head-10, AMT framework increases F1 score of ATL from 77.25 to 79.35. In the Tail-20, AFL's F1 score rises from 48.17 to 53.07. For the Tail-10, AMT framework boosts F1 scores of AFL and PEMSCL by 6.40 and 3.53, respectively.

6.2 Analyzing the Label Space Expansion

To verify the effectiveness of our proposed AMTL loss in mitigating performance degradation from label space expansion, we compare it with three other losses.

As shown in **Fig. 4**, the F1 scores of the four losses decrease as the label space expands. However, in most cases, the model with our AMTL loss consistently achieves higher F1 scores across different label spaces. For instance, when the size of label space is 20, the F1 score with AMTL is 78.10, compared to 77.06 with ATL; as the label space expands to 96, AMTL achieves an F1 score of 75.80, while ATL drops to 73.97. Notably, the advantage of AMTL becomes more pronounced in larger label spaces (e.g., with 80 and 96 labels). These results indicate that the AMTL *effectively mitigates performance degradation caused by label space expansion*. Additionally, we observe that compared to the other three losses, the performance degradation of AMTL loss is more gradual as the label space expands.

Loss	Head-10	Mid-76	Tail-20	Tail-10
ATL	77.25	67.15	44.90	40.76
with AMT	79.35	69.83	53.52	42.49
AML	76.51	65.60	43.60	35.29
with AMT	78.07	67.61	45.35	41.42
AFL	77.47	67.73	48.17	41.77
with AMT	79.14	69.85	53.07	48.17
Balanced Softmax	77.43	66.29	47.51	35.53
with AMT	77.74	67.54	46.80	38.27
PEMSCL	78.16	68.82	48.46	42.78
with AMT	78.84	69.94	50.73	46.31
SAT	77.36	67.24	46.48	38.27
with AMT	78.77	70.15	53.40	46.46
HingeABL	79.06	69.10	51.51	45.41
with AMT	79.24	69.86	48.99	45.13

Table 6: F1 results for the long-tail problem on the Re-DocRED dev set, using ATLOP as the representation module and employing $\text{BERT}_{\text{base}}$ for encoding.

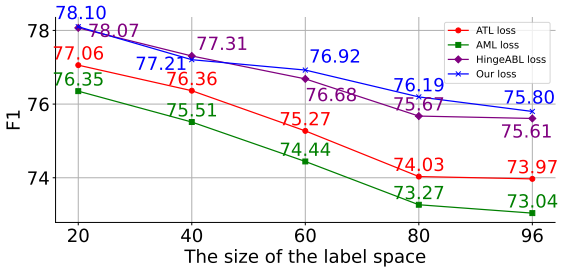

Figure 4: The effect of label space expansion on model performance is analyzed using different losses on the Re-DocRED dev set, with ATLOP as the representation module and $\text{BERT}_{\text{base}}$ for encoding.

6.3 Weakly Supervised Generalization Ability

Following Qi et al. (2024) and Gao et al. (2024), we evaluate the weakly supervised generalization of AMTL by training on the incompletely annotated DocRED dataset and testing on the more comprehensively annotated Re-DocRED dataset.

The results presented in **Table 7** indicate that AMTL significantly outperforms several competi-

Model	F1	Ign-F1
GAIN (Zeng et al., 2020) *	41.68	41.26
LogicRE-GAIN (Ru et al., 2021) *	41.89 (+0.21)	41.53 (+0.27)
MILR-GAIN (Fan et al., 2022) *	43.17 (+1.49)	42.89 (+1.63)
JMRL-GAIN (Qi et al., 2024) *	49.58 (+7.90)	47.85 (+6.59)
AMTL-GAIN (Ours)	**58.40** (+16.72)	**57.24** (+15.98)
ATLOP (Zhou et al., 2021) *	41.95	41.67
LogicRE-ATLOP (Ru et al., 2021) *	42.73 (+0.78)	42.47 (+0.80)
MILR-ATLOP (Fan et al., 2022) *	44.72 (+2.77)	44.30 (+2.63)
JMRL-ATLOP (Qi et al., 2024) *	47.54 (+5.59)	47.32 (+5.65)
AMTL-ATLOP (Ours)	**58.88** (+16.93)	**58.04** (+16.37)

Table 7: Weakly supervised generalization comparison results: training on DocRED and testing on Re-DocRED. Results with * are from Qi et al. (2024), and all methods employ $\text{BERT}_{\text{base}}$ for encoding.

Model	P	R	F1
DocGNRE (GPT) with $\text{RoBERTa}_{\text{large}}$			
DREEAM *	84.92	63.86	72.90
JMRL-DREEAM *	83.83 (-1.09)	65.92 (+2.06)	73.81 (+0.91)
AMTL-DREEAM	84.67 (-0.25)	68.07 (+4.21)	**75.47** (+2.57)
DocGNRE (mGPT) with $\text{BERT}_{\text{base}}$			
DREEAM *	81.71	58.23	68.00
JMRL-DREEAM *	82.55 (+0.84)	59.39 (+1.16)	69.08 (+1.08)
AMTL-DREEAM	79.78 (-1.93)	64.68 (+6.45)	**71.44** (+3.44)

Table 8: A comparison of our AMTL loss with plugin-based method JMRL, based on the DocGNRE test set results. Results marked with * are from Qi et al. (2024).

Loss	Training Time
ATL (Zhou et al., 2021)	40.04 minutes
AML (Wei and Li, 2022)	40.23 minutes
SSR-PU (Wang et al., 2022)	88.42 minutes
HingeABL (Wang et al., 2023)	40.13 minutes
AMTL (Ours)	41.91 minutes

Table 9: Comparison of training time for various losses using the ATLOP-backbone framework. All losses are trained for 30 epochs with a batch size of 4 on Re-DocRED dataset, using $\text{BERT}_{\text{base}}$ for encoding.

tive methods, including LogicRE (Ru et al., 2021), MILR (Fan et al., 2022), and JMRL (Qi et al., 2024). Specifically, AMTL-ATLOP achieves F1 and Ign-F1 scores of 58.88 and 58.04, respectively, marking improvements of 16.93 and 16.37 over the baseline ATLOP model. Furthermore, compared to the JMRL-ATLOP model, AMTL-ATLOP exhibits increases of 11.34 and 10.72 in F1 and Ign-F1 scores, respectively. These improvements suggest that AMTL not only possesses the ability to learn effectively from noisy data but also exhibits superior performance when confronted with datasets that feature more comprehensive and precise annotations. This further validates the strong generalization capability and robustness of AMTL in DocRE tasks.

6.4 Comparison with Plugin-based Methods

To evaluate the performance of our AMTL loss in comparison with different plugin-based methods, we integrate the AMTL loss and the competitive logical reasoning plugin JMRL (Qi et al., 2024) into the DREEAM (Ma et al., 2023) model for comparison.

The results in **Table 8** indicate that AMTL outperforms JMRL in terms of F1 score. In the AMTL-DREEAM configuration using $\text{BERT}_{\text{base}}$ for encoding, the F1 score improves by 3.44 compared to DREEAM, demonstrating a significant enhancement. Furthermore, AMTL-DREEAM improves F1 by 2.36 over JMRL-DREEAM. These results suggest that the AMTL loss exhibits a competitive advantage in overall performance.

6.5 Analyzing Cost

To verify the time cost of our loss, we compare the training times of different losses. As shown in **Table 9**, the training time of the AMTL loss under the ATLOP-backbone framework is 41.91 minutes, which is comparable to other losses (such as ATL, AML, and HingeABL). This indicates that our method achieves similar training time while maintaining good performance.

6.6 AMTL Loss vs. HingeABL Loss on Different Models

To further verify the effectiveness and advantages of our proposed AMTL loss, we conduct two sets of experiments. One investigates the performance of AMTL and the SOTA HingeABL loss across categories with varying frequencies, while the other evaluates how different models perform when trained with AMTL and HingeABL.

Performance Comparison Under Different Models. The experimental results in **Table 10** show that AMTL consistently outperforms HingeABL across all models and encoder settings ($\text{BERT}_{\text{base}}$ and $\text{RoBERTa}_{\text{large}}$). The improvements in F1 score range from 0.11 to 0.48, while Ign-F1 gains reach up to 0.70. These consistent gains demonstrate the effectiveness and robustness of the AMTL loss.

Long-Tail Problem Performance with Different Models. To further compare the performance of AMTL and HingeABL losses across different relation frequency categories, we conduct experiments using various backbone models, as shown in **Table 11**. AMTL consistently outperforms HingeABL,

Model	F1 (HingeABL)	F1 (AMTL)	Ign-F1 (HingeABL)	Ign-F1 (AMTL)
with $\text{BERT}_{\text{base}}$				
ATLOP	75.15	**75.63** (+0.48)	73.84	**74.44** (+0.60)
DocuNet	75.72	**76.01** (+0.29)	74.59	**74.73** (+0.14)
KD-DocRE	75.16	**75.63** (+0.47)	73.96	**74.53** (+0.57)
DREEAM	75.42	**75.71** (+0.29)	74.28	**74.98** (+0.70)
with $\text{RoBERTa}_{\text{large}}$				
ATLOP	79.79	**79.97** (+0.18)	78.82	**79.04** (+0.22)
DocuNet	79.43	**79.71** (+0.28)	78.39	**78.66** (+0.27)
KD-DocRE	79.34	**79.45** (+0.11)	78.26	**78.46** (+0.20)
DREEAM	79.85	**79.98** (+0.13)	78.80	**79.40** (+0.60)

Table 10: Performance comparison of AMTL loss and HingeABL loss on the Re-DocRED test set.

Model	Head-10	Mid-76	Tail-20	Tail-10
ATLOP				
+HingeABL	79.06	69.10	51.51	45.41
+AMTL	**79.24** (+0.18)	**69.86** (+0.76)	48.99 (-2.52)	45.13 (-0.28)
DocuNet				
+HingeABL	79.51	69.75	52.22	43.18
+AMTL	**80.14** (+0.63)	**70.10** (+0.35)	**54.82** (+2.60)	**51.81** (+8.63)
KD-DocRE				
+HingeABL	78.88	69.22	53.05	42.60
+AMTL	**79.68** (+0.80)	**69.97** (+0.75)	**54.67** (+1.62)	**46.49** (+3.89)

Table 11: Experimental results comparing AMTL loss and HingeABL loss on different base representation modules and four categories, using the F1 of $\text{BERT}_{\text{base}}$ on the Re-DocRED dev set.

especially on low-frequency categories (Tail-20 and Tail-10). For example, with DocuNet, AMTL improves F1 by 8.63 on Tail-10; with KD-DocRE, the gain is 3.89. When ATLOP is selected as the representation module, AMTL outperforms HingeABL on high-frequency categories, but it does not outperform HingeABL on two of the less frequent categories. These overall results further indicate that AMTL performs better on both high-frequency and less frequent categories.

We analyze the reasons as follows: While HingeABL reduces the impact of easily predictable negatives and emphasizes challenging minority classes, it still uses a single threshold and overlooks differences across relations. In contrast, our AMTL loss partitions the label space and learns adaptive thresholds for each sub-space, allowing finer adaptation to relation-specific characteristics, which better mitigates long-tail problem and avoids over-filtering.

6.7 Effect of Sub-label Space Number on Overall Threshold and Recall.

To further investigate the effect of the number n of sub-label space segments on the model, we conducted experiments summarized in **Table 12**. Intuitively, a larger n should benefit datasets with more labels, as it enables more precise control over the model's sensitivity to different types of relations.

Sub-label Space Number	2	3	4	5	6	7	8	9
Overall Threshold	10.68	11.74	12.24	12.58	12.71	12.91	13.03	13.08
Recall	78.31	73.24	72.85	70.49	69.69	69.05	69.12	68.79

Table 12: Effect of sub-label space number on overall threshold and recall.

However, experimental results reveal that increasing the number of sub-label space segments also raises the overall threshold, which is defined as the average of the thresholds across all subspaces. This increase in threshold subsequently leads to a decline in recall.

This is because each subspace's threshold is influenced not only by its own prediction distribution but also by those of other subspaces (see **Eq. 2**). As a result, finer label space partitioning tends to raise confidence thresholds for positive predictions, filtering out more low-confidence relations and thus lowering recall.

6.8 Analyzing Hyperparameters

To evaluate the impact of the hyperparameters (including the segment number n, the weighted average coefficient λ, and the label orderings) on performance, we conduct experimental analyses, with detailed results provided in **Appendix A**.

7 Conclusion

We propose a novel adaptive multi-threshold loss, AMTL, which effectively mitigates the performance degradation caused by label space expansion and the long-tail problem in DocRE tasks. AMTL first proposes to partition the label space into multiple segments and assign an adaptive threshold for each segment. Moreover, we design a multi-threshold framework that enables the collaborative application of multiple losses across different label space segments, which outperforms the single-threshold methods used in prior work. Experiments show that AMTL significantly improves the predictive performance of various models and achieves SOTA results on four datasets, further demonstrating its superiority. AMTL also exhibits effective generalization capabilities, performing well on both partially labeled and fully labeled datasets. Since our method is independent of specific models, it holds potential for wide applicability in other multi-label classification tasks.

Limitations

Despite our AMTL loss and AMT framework demonstrating advantages in the DocRE task, there are still some limitations. Firstly, although our AMT framework effectively combines different losses and maintains stable performance, identifying a clear pattern for the optimal loss combination remains challenging, and we plan to investigate this further in future work. Moreover, as demonstrated in our experiments, the AMTL loss demonstrates more effective than other losses in mitigating performance degradation caused by label space expansion. However, when the number of relation types increases to 96, our model experiences a certain degree of performance decline compared to its performance with 20 relation types. This raises concerns about its scalability to real-world scenarios, where the number of relation types may be significantly larger. The performance of our approach on such scenarios remains unexplored.

Acknowledgments

The authors sincerely thank the reviewers for their valuable comments, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057).

References

Shulin Cao, Jiaxin Shi, Zijun Yao, Xin Lv, Jifan Yu, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jinghui Xiao. 2022. Program transfer for answering complex questions over knowledge bases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8128–8140.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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) (NAACL-HLT)*, pages 4171–4186.

Shengda Fan, Shasha Mo, and Jianwei Niu. 2022. Boosting document-level relation extraction by mining and injecting logical rules. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10311–10323.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. TTM-RE: memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2598–2609.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5495–5505.

Wanlong Liu, Li Zhou, Dingyi Zeng, Yichen Xiao, Shaohuan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. Beyond single-event extraction: Towards efficient document-level multi-event argument extraction. In *Findings of the Association for Computational Linguistics, (ACL)*, pages 9470–9487.

Yichun Liu, Zizhong Zhu, Xiaowang Zhang, Zhiyong Feng, Daoqi Chen, and Yaxin Li. 2023. Document-level relationship extraction by bidirectional constraints of beta rules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2256–2266.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15453–15464.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. Dreeam: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1971–1983.

Kunxun Qi, Jianfeng Du, and Hai Wan. 2024. End-to-end learning of logical rules for enhancing document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pages 7247–7263.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1239–1250.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8472–8487.

Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. 2023. Adaptive hinge balance loss for document-level relation extraction. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3872–3878.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4123–4135.

Ye Wang, Huazheng Pan, Tao Zhang, Wen Wu, and Wenxin Hu. 2024. A positive-unlabeled metric learning framework for document-level relation extraction with incomplete labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 19197–19205.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2000–2008.

Xiaolong Xu, Chenbin Li, Haolong Xiang, Lianyong Qi, Xuyun Zhang, and Wanchun Dou. 2024. Attention based document-level relation extraction with none class ranking loss. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 764–777.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–17.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640.

Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, and Yongxue Wu. 2024. SRF: enhancing document-level relation extraction with a novel secondary reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15426–15439.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3999–4006.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4538–4544.

A Hyperparameter Analysis

A.1 Analyzing Sub-label Space Segments

To evaluate the impact of the number n of sub-label space segments on performance, we partition the label space into 9 segments. **Fig. 5** illustrates the F1 score variations for our proposed loss, AMTL, as the number of sub-label space segments changes. When the number of segments is 1, AMTL can be regarded as equivalent to ATL. As the number of segments increases beyond 1, we insert a new segment for every ten labels. For example, with 2 segments, one is inserted before the first label and another after the tenth label.

The F1 score peaks at 75.80 when the number of sub-label space segments is 4. However, as the number of segments increases beyond 4, the F1 score begins to decline, eventually dropping to 75.06 with 9 segments. This suggests that an excessive number of segments may reduce the model's performance, likely due to more scores of predicted relations exceeding the threshold, resulting in an increased number of positive classes.

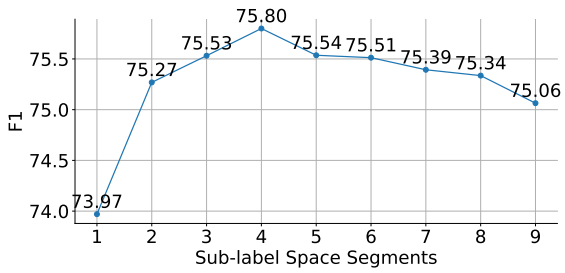

Figure 5: Performance with different numbers of sub-label space segments, evaluated on the Re-DocRED dev set, using ATLOP as the representation module and $\text{BERT}_{\text{base}}$ for encoding.

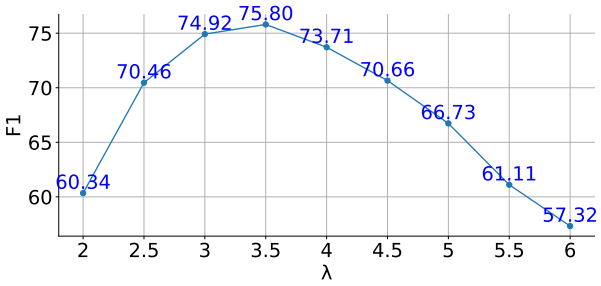

Figure 6: Performance varies with changes in the λ value, evaluated on the Re-DocRED dev set, using ATLOP as the representation module and $\text{BERT}_{\text{base}}$ for encoding.

A.2 Effect of λ in AMTL

As shown in **Fig. 6**, the model's performance exhibits a clear fluctuation with changes in the λ value. When λ is set to 3.5, the model achieves the optimal F1 score of 75.80 on the Re-DocRED dev set. In contrast, when the λ value is too high or too low, the performance of the model significantly deteriorates. For instance, at $\lambda = 6$, the F1 score drops to 57.32. This indicates that the λ value has a significant impact on the model's performance, and selecting an appropriate λ is crucial for achieving the best results.

A.3 Effect of Label Ordering in AMTL

Table 13 compares performance under three label ordering strategies: our proposed ordering, the original ordering, and random ordering, on the Re-DocRED dev set.

As shown in **Table 13**, the model with AMTL using our proposed label ordering achieves the highest F1 score of 75.80 and Ign-F1 of 74.59, outperforming the other configurations. Specifically, the model with the original label ordering achieves a slightly lower F1 score of 75.51 and Ign-F1 of 74.27, while the random ordering results in a similar performance with an F1 score of 75.48 and Ign-F1 of 74.34. These results highlight the importance of label ordering in improving model performance.

Method on Re-DocRED Dev Set	F1	Ign-F1
ATLOP	73.93	73.04
ATLOP with AMTL (Our Ordering)	**75.80**	**74.59**
ATLOP with AMTL (Original Ordering)	75.51	74.27
ATLOP with AMTL (Random Ordering)	75.48	74.34

Table 13: Results with different label ordering.

B Detailed Description of the Datasets

DocRED (Yao et al., 2019), is a large-scale, manually annotated dataset constructed from Wikipedia, and it is one of the largest datasets in DocRE. The dataset consists of 5,053 documents, with 3,053 used for training, and 1,000 each for development and testing.

DWIE (Zaporojets et al., 2021) is an entity-centric dataset that contains four natural language processing subtasks. To ensure consistency, we follow the method of Ru et al. (2021) to process the original DWIE dataset, resulting in 602 documents for training, 98 for development, and 99 for testing.

Re-DocRED (Tan et al., 2022b) is built on the DocRED (Yao et al., 2019) dataset, addressing the annotation errors and gaps of DocRED, and offering more accurate and comprehensive annotations. The dataset includes 3,053 documents for training, 500 for development, and 500 for testing.

DocGNRE (Li et al., 2023) is a dataset that uses the powerful generative capabilities of ChatGPT to expand and enhance the Re-DocRED dataset. It includes two training sets (GPT set and mGPT set) and a test set. The test set is generated through distant supervision using ChatGPT and has undergone rigorous manual verification to ensure high quality and reliability, making it suitable for accurate model evaluation. Each of the two training sets contains 3,053 documents, while the test set contains 500 documents.