附件 1

南开大学推荐免试攻读研究生申请表暨诚信承诺书

学号	2312141	姓名	张德民	性别	男
民族	汉	出生日期	2004 年 11 月 24 日	政治面貌	共青团员
学院	密码与网络空间安全学院		专业	信息安全	
手机号码	18339765000		电子邮箱	2312141@mail.nankai.edu.cn	
申请类型（单选）	☑普通类推免名额　　□支教团推免名额				
本人与学院遴选指标相关的材料列表	(学生可将本人与**本学院遴选指标**相关的材料列在下方，如科研成果、竞赛获奖、科研训练表现、参军入伍服兵役、志愿服务、国际组织实习等) 1.ICDAR 2025 论文，作者排序第一，CCF-C 类 2.志愿服务时长 81.5 小时				

本人充分了解并理解《南开大学推荐优秀应届本科毕业生免试攻读研究生工作实施办法》（南党发〔2021〕63 号）和所在学院《推免细则》，并郑重承诺：

1.本人所提供的一切材料真实可靠，无弄虚作假，如有不实之处，本人愿意承担由此造成的一切后果；

2.如获得推免生资格，本人将慎重填报志愿，绝不浪费学校推免名额。

承诺人：

年　　月　　日

（本申请书必须由申请人亲笔签名方为有效，交所在学院本科教学工作办公室备查）

综合素质评价审查材料目录（不包含公能素质部分）

学院：密码与网络空间安全学院　专业：信息安全　　学号：2312141　　姓名：张德民

请根据个人实际情况填写此目录（其他项目可自行添加行），并按填写顺序排序所有证明材料后附证明材料要求真实、直观。

此目录纸张横向单面打印、、附在“承诺书”后

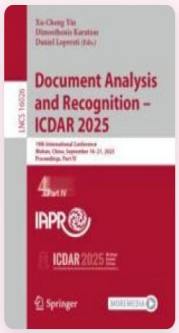
		加分项目 ① 赛事获奖请直接直接填写右边各列 ② 论文加分请写明论文题目后填写右边各列 ③ 辅修专业课程请写明辅修专业 ④ 参军入伍及系列情况写“参军入伍” 每类事项行数不够请在该类目下自行添加	此处填写获奖等级： ① 赛事获奖请填写“一等奖/二等奖/三等奖/金奖/银奖/铜奖” ② 论文发表请填写“发表期刊-类别”（eg. “IEEE TDSC”-CCF 英文 A 类） ③ 辅修专业课程总学分数 ④ 参军入伍/嘉奖/优秀士兵/三等功及以上	证明材料类别： ① 赛事获奖证书/获奖公告页面截图（材料里需标注出个人获奖信息所在行） ② 录用邮件页面截图/期刊网页截图+论文原文文本（证明自己是一作） ③ 辅修成绩单
特殊学术专长	重要赛事	高教社杯全国大学生数学建模竞赛本科组		
		全国大学生电子设计竞赛本科组		
		“挑战杯”全国大学生课外学术科技作品竞赛		
		中国国际大学生创新大赛（原中国国际“互联网+”大学生创新创业大赛）		
科研训练	各专业代表性赛事	ACM-ICPC 亚洲区域赛		
		全国大学生信息安全竞赛		
		全国密码技术竞赛		
		全国大学生计算机系统能力大赛		

		全国大学生物联网设计竞赛		
	一般赛事	全国大学生数学建模竞赛天津赛区		
		天津市“挑战杯”大学生课外学术科技作品竞赛		
		中国国际大学生创新大赛天津赛区（原中国国际“互联网+”大学生创新创业大赛天津赛区）		
		天津市物联网大赛		
		京津冀大学生信息安全网络攻防大赛（原天津市大学生信息安全网络攻防大赛）		
	辅修			
特殊学术专长	论文发表	Class-Agnostic Region-of-Interest Matching in Document Images	ICDAR 2025-CCF 英文 C 类	录用邮件页面截图/期刊网页截图+论文原文文本
	论文发表			
	论文发表			
参军入伍				
其他材料				

Class-Agnostic Region-of-Interest Matching in Document Images

Conference paper | First Online: 16 September 2025

pp 446–464 | Cite this conference paper



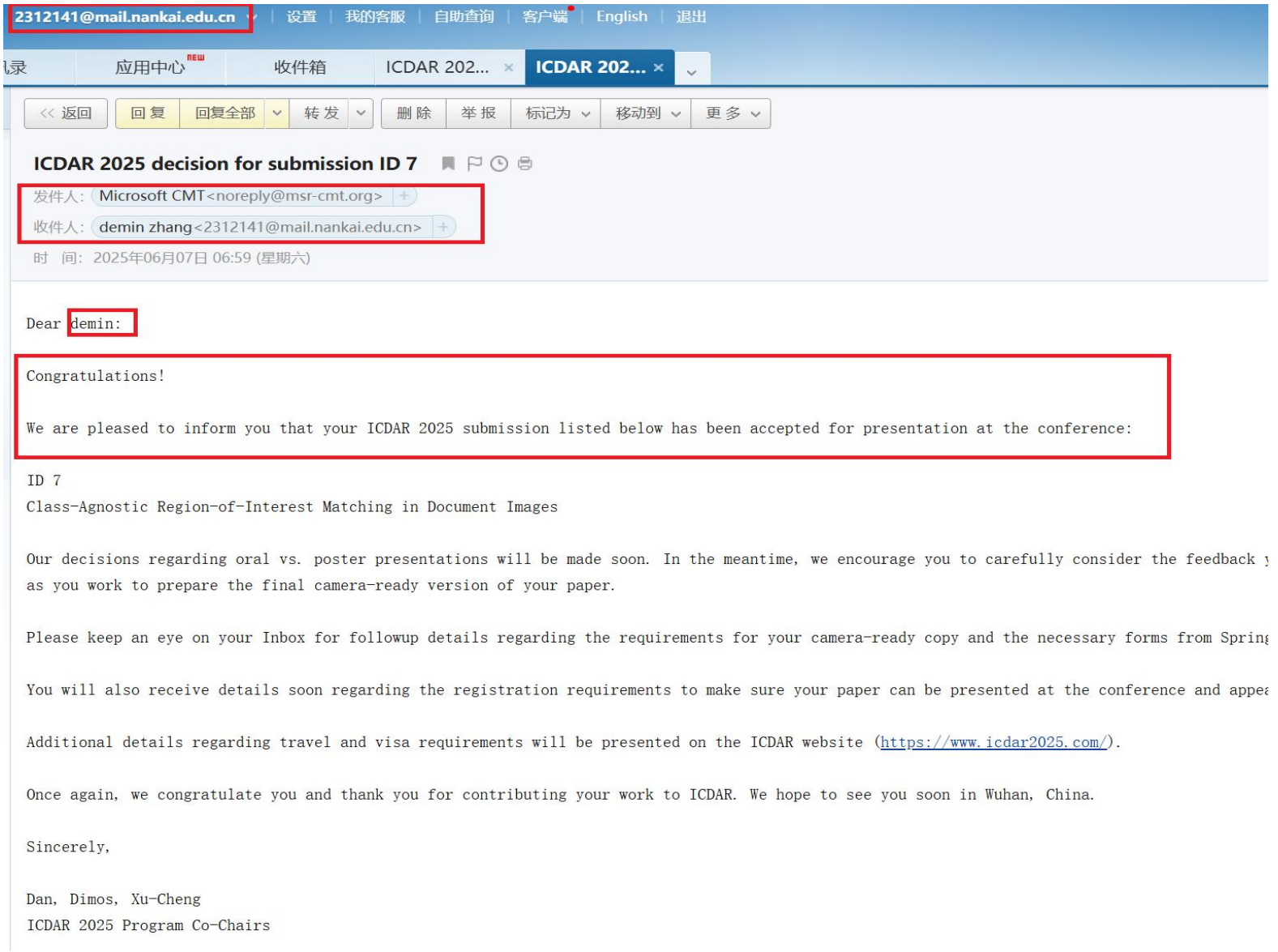
Document Analysis and Recognition – ICDAR 2025

(ICDAR 2025)

Demin Zhang, Jiahao Lyu, Zhijie Shen & Yu Zhou

Part of the book series: Lecture Notes in Computer Science ((LNCS,volume 16026))

Included in the following conference series:
International Conference on Document Analysis and Recognition


2312141@mail.nankai.edu.cn
设置
我的客服
自助查询
客户端
English
退出
应用中心
收件箱
ICDAR 202...
ICDAR 202...
<< 返回
回 复
回复全部
转 发
删 除
举 报
标记为
移动到
更 多
ICDAR 2025 decision for submission ID 7
发件人: Microsoft CMT<noreply@msr-cmt.org>
收件人: demin zhang<2312141@mail.nankai.edu.cn>
时 间: 2025年06月07日 06:59 (星期六)
Dear demin:
Congratulations!
We are pleased to inform you that your ICDAR 2025 submission listed below has been accepted for presentation at the conference:
ID 7
Class-Agnostic Region-of-Interest Matching in Document Images
Our decisions regarding oral vs. poster presentations will be made soon. In the meantime, we encourage you to carefully consider the feedback
as you work to prepare the final camera-ready version of your paper.
Please keep an eye on your Inbox for followup details regarding the requirements for your camera-ready copy and the necessary forms from Spring
You will also receive details soon regarding the registration requirements to make sure your paper can be presented at the conference and appea
Additional details regarding travel and visa requirements will be presented on the ICDAR website (https://www.icdar2025.com/).
Once again, we congratulate you and thank you for contributing your work to ICDAR. We hope to see you soon in Wuhan, China.
Sincerely,
Dan, Dimos, Xu-Cheng
ICDAR 2025 Program Co-Chairs

Class-Agnostic Region-of-Interest Matching in Document Images

Demin Zhang[1[0009−0006−0069−5190]*], Jiahao Lyu[2,3[0000−0003−2051−8045]*], Zhijie Shen[4[0009−0008−6135−7958]], and Yu Zhou[1[0000−0003−4188−9953]✉]

[1] VCIP & TMCC & DISSec, College of Computer Science & College of Cryptology and Cyber Science, Nankai University, China
2312141@mail.nankai.edu.cn, yzhou@nankai.edu.cn
[2] Institute of Information Engineering, Chinese Academy of Science, China
[3] School of Cyber Security, University of Chinese Academy of Science, China
lvjiahao@iie.ac.cn
[4] Department of Computer Science and Technology, Tsinghua University, China
shenzj24@mails.tsinghua.edu.cn

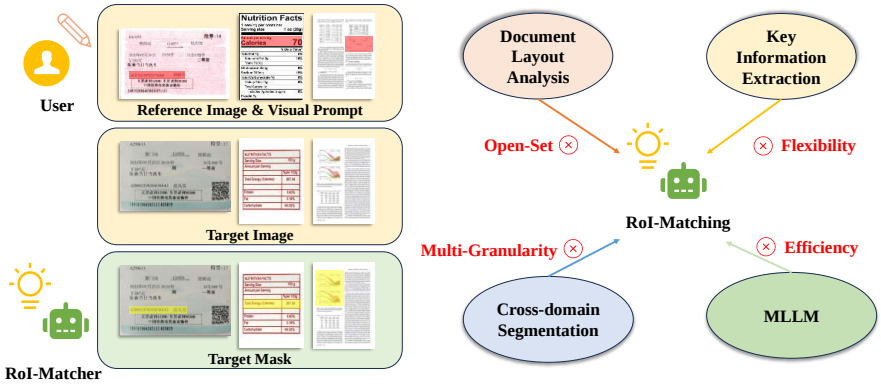
Abstract. Document understanding and analysis have received a lot of attention due to their widespread application. However, existing document analysis solutions, such as document layout analysis and key information extraction, are only suitable for fixed category definitions and granularities, and cannot achieve flexible applications customized by users. Therefore, this paper defines a new task named "Class-Agnostic Region-of-Interest Matching" ("RoI-Matching" for short), which aims to match the customized regions in a flexible, efficient, multi-granularity, and open-set manner. The visual prompt of the reference document and target document images are fed into our model, while the output is the corresponding bounding boxes in the target document images. To meet the above requirements, we construct a benchmark RoI-Matching-Bench, which sets three levels of difficulties following real-world conditions, and propose the macro and micro metrics to evaluate. Furthermore, we also propose a new framework RoI-Matcher, which employs a siamese network to extract multi-level features both in the reference and target domains, and cross-attention layers to integrate and align similar semantics in different domains. Experiments show that our method with a simple procedure is effective on RoI-Matching-Bench, and serves as the baseline for further research. The code is available at https://github.com/pd162/RoI-Matching.

Keywords: RoI-Matching · Document Analysis · Benchmark.

1 Introduction

Document images play a vital role in storing and transmitting information, making content and structure understanding a key challenge. As a result, tasks

* Equal Contribution. ✉ Corresponding Author.



(a) Pipeline of RoI-Matching **(b) Difference between Other Tasks**

Fig. 1.1. The task definition of RoI-Matching. (a) shows the pipeline of RoI-Matching. Given some document images by the user, RoI-Matcher responds to the corresponding mask, even if the user only provides visual prompt without class labels ("name", "number of calories", and "figure" from left to right). (b) demonstrates the difference of RoI-Matching between other Tasks.

like processing [1, 2], detection [3–5], recognition [6–10], spotting [11–13], retrieval [14, 15], and understanding [16, 17] have gained attention for their broad applicability. To advance document understanding, various downstream tasks have been developed. For example, key information extraction identifies vital content, while layout analysis locates important document components. These tasks support real-world applications such as automating processes in legal and financial domains.

Despite significant advancements in downstream tasks, challenges remain in effectively matching user-customized regions of interest (RoIs) in documents, which is essential for practical applications in smart offices and ensuring user-friendly and flexible interactions. This difficulty arises because customized RoIs cannot be captured by a limited set of predefined concepts. Current document layout analysis methods are constrained by fixed categories, which hinder their generalization in open-world settings. As documents continue to diversify and evolve, developing more robust methods for efficient RoI matching becomes crucial, particularly in the presence of complex and unstructured content.

In this paper, we propose a novel task, class-agnostic region-of-interest matching (RoI-Matching), aimed at efficiently matching user-defined regions of interest in document images. RoI-Matching is designed to match customized regions in a **flexible, efficient, multi-granular, and open-set** manner. **"Flexible"** refers to the diversity of user inputs, which are not restricted by language descriptions. To accommodate this flexibility, we introduce the concept of a visual prompt. Given the flexibility of the mask, we use a reference mask corresponding to the

reference image to represent the customized visual prompt. **"Efficient"** indicates that the RoI-Matching is designed for high inference speed, ensuring its practical application. **"Multi-granular"** implies that the model must handle a range of scenarios, from word-level to line-level and paragraph-level instances. **"Open-set"** denotes the absence of class limitations in this task, ensuring its ability to generalize. The general inputs and outputs are illustrated in Fig. 1.1 (a).

RoI-Matching differs from other related tasks, as shown in Fig. 1.1(b). First, document layout analysis [18–20, 10] detects and segments different types of regions and analyzes their relationships within the document image. However, it is limited to fixed categories and cannot handle open-set scenarios. Key information extraction [21–23] involves mining and interpreting multi-modal (semantic, layout, and visual) cues from visually rich document images to extract text information for specified categories. However, this process lacks flexibility, as it is only applicable to structured visually rich documents. Few-shot cross-domain segmentation [24–27] aims to learn a model that can perform segmentation on novel classes with only a few pixel-level annotated images. Although this is an open-set task, it is unsuitable for multi-granularity scenarios within document environments. In recent years, Multi-modal Large Language Model (MLLM) [28, 29] solves many challenging problems of document images in an end-to-end manner. But the low inference speed of MLLMs fails to meet the efficiency requirements due to the limitations of auto-regressive decoding. Furthermore, studies have shown that MLLMs perform poorly in localization tasks. Thus, the introduction of RoI-Matching and the construction of the related dataset are meaningful for further enabling flexible and efficient understanding of document images.

To meet the above requirements, we construct a benchmark for the RoI-Matching task and propose a simple yet effective baseline, RoI-Matcher. Specifically, we curate a new dataset, RoI-Matching-Bench, designed for region-of-interest matching in documents, enabling performance evaluation under real-world conditions. The dataset covers diverse document types, from standardized forms to complex unstructured layouts, to assess generalization.We also introduce RoI-Matcher, a simple yet efficient method that uses a Siamese network to extract multi-granularity features from reference and target images in parallel. Two cross-attention layers are used to project reference regions into the target domain, while additional designs mitigate overlapping response issues common in document images. Our main contributions are threefold:

- We define a new task, RoI-Matching, to match the customized regions in a flexible, efficient, multi-granular, and open-set manner.
- We also construct the corresponding RoI-Matching-Bench with three levels of difficulties, which allows researchers to evaluate the performance of RoI-Matching approach under real-world conditions. Two evaluation metrics are also introduced in the macro and micro perspectives.
- We propose a new framework RoI-Matcher, which employs a siamese network to extract multi-level features both in the reference and target domains, and

cross-attention layers to integrate and align similar semantics in different domains. Some additional designs also mitigate the impact of overlapping responses and improve performances.

2 Related Works

2.1 Document Layout Analysis

Document layout analysis (DLA) methods fall into uni-modal and multi-modal categories. Uni-modal approaches focus mainly on visual features to analyze layouts, often adapting object detection and instance segmentation techniques. For example, PubLayNet [18] employs Faster R-CNN [30] and Mask R-CNN [31] for layout detection. TransDLANet [32] utilizes three parameter-shared MLPs atop ISTR, while SwinDocSegmenter [33] combines high- and low-level image features to initialize queries in DINO [34]. SelfDocSeg [35] pretrains its image encoder with pseudo-layouts using the BYOL [36] framework before fine-tuning on layout datasets.

While these approaches improve model performance, they remain limited in fully leveraging document semantics. Recently, multi-modal methods have been introduced to address complex DLA challenges. For example, MFCN [19] employs skip-gram to capture sentence-level textual features and integrates them with visual features in the decoder. VSR [20] builds a full text embedding map combining Char-grid, word-grid, and sentence-grid levels, using dual backbones to extract and fuse visual and textual features via a multi-scale-adaptive-aggregation module. Notable models such as LayoutLM [37–39], BEiT [40], DiT [41], UDoc [42], and StrucText [23, 43] adopt a pre-training then fine-tuning approach. Extensive pre-training on large document datasets enables these models to excel in downstream tasks like image classification, layout analysis, and information extraction.

2.2 Cross-domain few-shot Segmentation

Cross-domain few-shot semantic segmentation (CD-FSS) targets segmenting unseen classes in query images using only a few annotated samples. PATNet [24] pioneered this task by designing a feature transformation module to convert domain-specific features into domain-agnostic ones. RestNet [26] enhances pre-transformation features with a lightweight attention module and merges post-transformation features via residual connections to preserve key domain information. RD [25] uses a memory bank to restore source domain meta-knowledge and augment target data. APSeg [44] adapts SAM to CD-FSS with a dual prototype anchor transformation (DPAT) module and a meta prompt generator (MPG) for efficient learning. ABCDFSS [27] introduces a consistency-based contrastive learning scheme that estimates attached layer parameters without overfitting and yields domain-shift robust masks. Given their similar input-output settings, we compare our method against these approaches.

2.3 Key Information Extraction

Currently, methods for Key Information Extraction (KIE) can be divided into two main types: OCR-dependent and OCR-free models. OCR-dependent models use optical character recognition (OCR) to extract textual information. Early KIE methods [45, 46] construct layout-aware or graph-based representations for KIE tasks with sequence labeling of OCR inputs. Nevertheless, these methods rely on text with a proper reading order or need additional modules for OCR serialization, which is not always feasible where the document layout might be intricate or unordered in real-world scenarios. To tackle the serialization problem, certain methods [22, 47] make use of extra detection or linking modules to model the intricate connections between text blocks or tokens. Additionally, methods based on generation [48, 21, 49] have been put forward to reduce the effort of post-processing and the need for task-specific link designs. Another category of OCR-free methods [50, 51] offers an alternative, which either utilize OCR-aware pre-training or incorporates OCR modules within an end-to-end framework. Donut[52], Omniparser[53] and other methods adopt a text reading pretraining objective and generate structured outputs consisting of text and entity token.

3 Benchmark

3.1 Statistics and Visualizations

To build the RoI-Matching-Bench benchmark for the RoI-Matching task, we leverage six existing document datasets spanning key information extraction, document layout analysis, and more. Dataset details and sources are listed in Table 3.1. The task is divided into three difficulty levels—Level I, II, and III—based on matching granularity and complexity. For Levels I and III, masks are generated directly from open-source box annotations. For Level II, we first define a semantic category schema, followed by manual annotation from trained annotators. All annotations undergo strict visual verification to ensure accuracy and consistency. We group semantically and visually similar images into categories, pair them, and randomly sample to balance the distribution and avoid outliers. Categories with distinctive features (e.g., bank card numbers, forms) are selected along with corresponding images, while content lacking significant visual features, such as article paragraphs, is excluded. For datasets with predefined splits, we use the original partitions; for those without validation sets, we create stratified random subsets from the training data as validation sets. Definitions of the three levels are provided below.

Level I. As the simplest and most common case, we define Level I as Reference-Target pairs with similar visual elements and positions. This level of matching is commonly found in scenarios like standardized forms, including invoices, ID cards, tickets, etc. To construct the dataset following the above definition, we regard ICDAR2023-SVRD [54] as baseline benchmark and automatically create an organized benchmark, due to its visually similar images and detailed annotations.

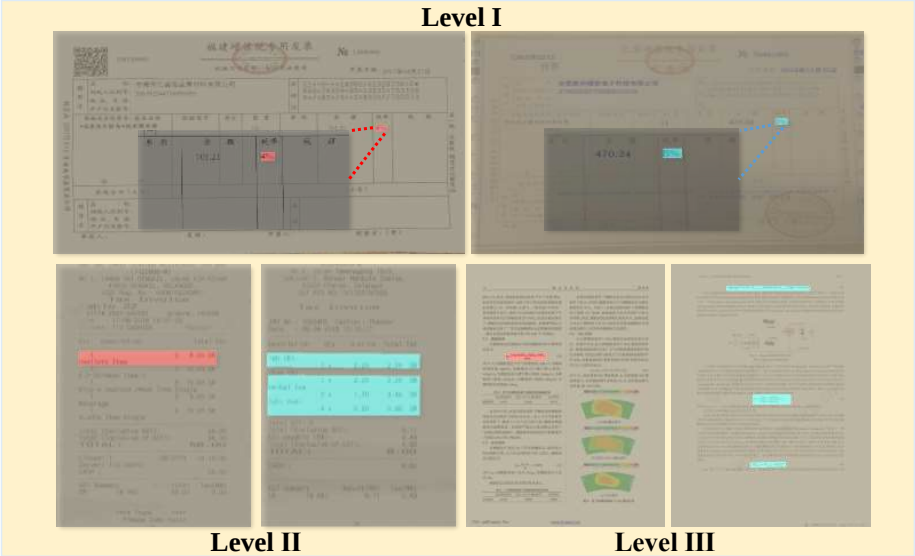


Fig. 3.1. The visualizations of RoI-Matching-bench. The reference image (left) and the target image (right) are image pairs. Hint: these visual prompts indicate semantic labels, which are "Tax Ratio", "Items in Receipt", "Equation".

Table 3.1. Derivation, scenarios, and statistics on the benchmark of RoI-Matching.

Difficulties	Derivation	Images	Pairs			Examples of Scenario
			Train	Val	Test	
Level I	SVRD [54]	660	107k	253	1260	Invoice, ID Card, Ticket.
Level II	POIE [55], SROIE [56]	233	13k	851	913	Receipt, Nutrition Fact.
Level III	D4LA [57], M6Doc [32], etc.	4312	33k	809	673	Paper, Advertisement.

Level II. Compared to Level I, Level II includes reference-target pairs that are visually similar but differ in position. We use two datasets from the field of key information extraction (KIE) to construct this sub-benchmark. To ensure consistency with other sub-benchmarks, we manually sample and annotate these data.

Level III. As the most challenging sub-task, Level III consists of the samples which are significantly different in both visual appearance and position. These samples are commonly found in advertisements and academic papers. We use three datasets from the field of document layout analysis and select classes that contain visual elements with distinct features and similar semantics, such as equations, figures, and titles of academic papers.

Furthermore, we gather statistics of RoI-Matching-Bench and visualize the position distribution in Fig. 3.1 and Fig. 3.2, which demonstrate the construction of our benchmark across multi-grained data, satisfying the generalized rule.

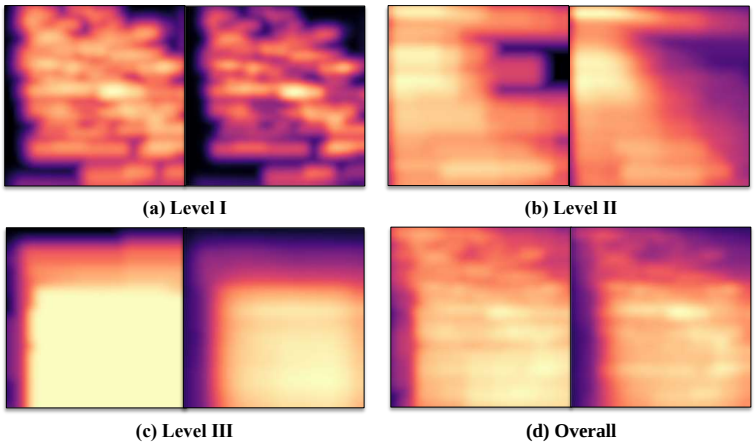


Fig. 3.2. The position distribution of RoI-Matching benchmark. There are two images in each group, the left is the visualization of the reference images, and the right is the target images.

3.2 Evaluation Protocols

We report two metrics: one from the pixel-wise protocol for common semantic segmentation and another from the instance-wise protocol for scene text detection.

Mean Interaction-of-Union (mIoU). Inspired by [58], we use mIoU to evaluate the effectiveness of our method from the pixel-level perspective. The formulation of mIoU is shown as Eq. (1). Let p_{ij} represent the number of pixels of class i predicted to belong to class j, where there are k different classes. $k = 1$ in this paper due to the setting of class-agnostic.

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{1}$$

F-Measure. F-measure is widely used to evaluate binary classifiers in machine learning. In our evaluation process, we first convert the final binary mask into instance-level prediction through connected component analysis. Then, we calculate IoU score between the ground truth and the prediction. We consider the samples with $IoU > 0.5$ as correctly predicted. Following standard practices in scene text detection, we use precision, recall, and F-measure to evaluate the RoI-Matching results.

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

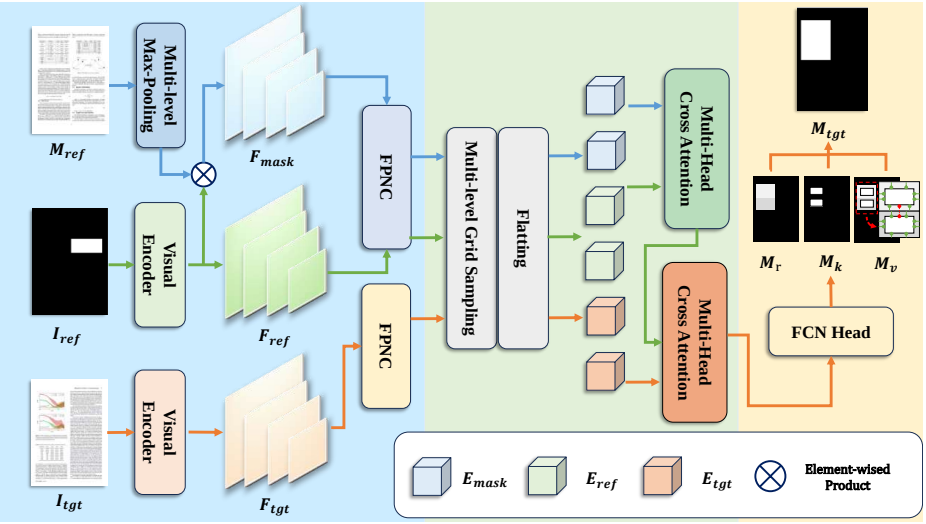


Fig. 4.1. The architecture of RoI-Matcher. The background colors represent three stages of our model, blue for Visual Perception, green for Cross-granularity Attention, and yellow for Segmentation Head, respectively. The colors of arrows indicate flows of different inputs, blue for the reference mask, green for the reference image, and red for the target image.

4 Method

In this section, we introduce the class-agnostic region-of-interest matching task first. Then an effective baseline for our benchmark is described in detail. The baseline consists of three main components: Visual Perception, Cross-granularity Attention, and Segmentation Head. Furthermore, we introduce task-aware label generation and optimization.

4.1 Task Definition

In this section, we first propose the class-agnostic region-of-interest matching task in document scenarios. This task can be formalized as follows. The prior are reference document image $I_{ref} \in \mathbb{R}^{H_{ref} \times W_{ref} \times 3}$ and user-customized visual prompt, representing a binary mask $M_{ref} \in \mathbb{R}^{H \times W}$. Because a mask can represent any form of visual prompt. During the inference stage, given the target document image $I_{tgt} \in \mathbb{R}^{H_{tgt} \times W_{tgt} \times 3}$, the model $\mathcal{M}$ aims to find the region of interest, where a binary mask $M_{tgt} \in \mathbb{R}^{H_{tgt} \times W_{tgt}}$ for I_{tgt} represents the final result. Note that there is no category information during the whole pipeline. Therefore, we define this task as the class-agnostic region-of-interest matching.

4.2 Visual Perception

Visual Perception aims to extract multi-level features F_{ref} and F_{tgt} from reference image I_{ref} and target image I_{tgt}. The extraction is formulated by Eq. (3), where k is the index of $\{ref, tgt\}$.

$$F_k = VE(I_k) \in \{\mathbb{R}^{\frac{W_k}{4} \times \frac{W_k}{4} \times C}, \mathbb{R}^{\frac{W_k}{8} \times \frac{W_k}{8} \times 2C}, \mathbb{R}^{\frac{W_k}{16} \times \frac{W_k}{16} \times 4C}, \mathbb{R}^{\frac{W_k}{32} \times \frac{W_k}{32} \times 8C}\}. \quad (3)$$

Moreover, the reference mask provided by user M_{ref} interacts initially with F_{ref} level by level, as conducted by Eq. (4). $MP(x, y)$ indicates the Max Pooling operator, where x is feature map and y is the sampling ratio. $\otimes$ is the element-wise dot production. i represents i-th layer.

$$F^i_{mask} = F^i_{ref} \otimes MP(M^i_{ref}, 2^i). \quad (4)$$

Then four feature maps are aggregated using the efficient FPNC [59] respectively. FPNC is a kind of FPN variant and more efficient than classical FPN. This process is formulated by Eq. (5), where j is the index of $\{ref, tgt, mask\}$. Note that $\mathcal{F}_{mask}$ and $\mathcal{F}_{ref}$ is generated by weight-shared FPNC, because these two feature maps are homologous.

$$\mathcal{F}_j = FPNC(F_j). \quad (5)$$

4.3 Cross-granularity Attention

After extracting from the reference and target document image, the key problem is how to transfer from the reference visual prompt to the target domain. Considering the high-resolution of the document image, we first sample the multi-level features into smaller feature maps to alleviate the computational burden. Then the sampled features are flattened into different embeddings, termed as $E_{mask}, E_{ref}, E_{tgt}$. The grid sampling and flattening process is written as Eq. (6).

$$E_j = Flatten(GS(\mathcal{F}_j)) \in \mathbb{R}^{\frac{H_k W_k}{256} \times 4C} \quad (6)$$

Extracted by multi-level feature maps, $E_{mask}, E_{ref}, E_{tgt}$ contains rich vision information from different granularities. The interaction between different granularities and domains is realized by Multi-Head Cross Attention [60]. The detailed process is formulated as Eq. (7) and Eq. (8). For the first cross attention, we regard the E_{mask} as the query and E_{ref} as the key and value. Two cross-attention operations ensure transfering the visual prompt from the reference domain to the target domain, preventing information loss.

$$E_r = MHCA(Q = E_{mask}; K = E_{ref}; V = E_{ref}), \quad (7)$$

$$E_{total} = MHCA(Q = E_{tgt}; K = E_r; V = E_r). \quad (8)$$

4.4 Segmentation Head

After interacting features from different domains and granularities, we construct the Segmentation Head to generate the final mask $\hat{M}_{tgt}$. Considering the overlapped regions (as shown in M_r in Fig. 4.1) in document images, we introduce the subtle mask representation inspired by PAN [61]. Compared with a single binary mask, we use the 6-channel mask to represent $\hat{M}_{tgt}$. Naively, we adopt Fully Convolution Network (FCN) [58] to predict 6 channels. The first channel is regions M_r, and the second is kernel M_k, where its label generation will be described in Section 4.5. The last four channels indicate similarity vectors, which guide the pixels in the text area to the corresponding kernel. The detailed post-processing follows PAN [61], where the M_k decides the number of regions, then M_v assists in refining the region more precisely.

$$M'_{tgt} = [M_r, M_k, M_v] = UpSampling(FCN(E_{total})) \in \mathbb{R}^{6\times H_t \times W_t}. \quad (9)$$

4.5 Label Generation

Due to the dense information in document images, the regions of interest could be overlapped. Therefore, we need to use the strategy during the label generation stage. The label generation for the target mask M_k is inspired by PSENet [62]. Specifically, given a target image with N_r regions of interest, each region can be described by a set of segments, where n is the number of vertexes. In this task, user-customized visual prompt is represented by a binary mask, which is regarded as an irregular polygon. So n is not fixed.

$$G = \{S_k\}_{k=1}^{n}. \quad (10)$$

The shrink offset D of shrinking is computed from the perimeter L and area A of the original polygon: where r is the shrink ratio, set to 0.4 empirically. The final label M_k is generated by the original label G and offset D, using the Vatti clipping algorithm [63].

$$D = \frac{A(1-r^2)}{L}. \quad (11)$$

4.6 Optimization

Followed by PAN [61], our optimization target can be formulated as:

$$\mathcal{L} = \mathcal{L}_{region} + \alpha\mathcal{L}_{kernel} + \beta(\mathcal{L}_{agg} + \mathcal{L}_{dis}), \quad (12)$$

where $\mathcal{L}_{region}$ is the loss of regions and $\mathcal{L}_{kernel}$ is the loss of kernels. α and β are the balanced factors, we set them to 0.5 and 0.25 respectively.

Considering segmentation task, we adopt dice loss [64] to supervise $\hat{M}_r$ and $\hat{M}_k$. $\mathcal{L}_{region}$ and $\mathcal{L}_{kernel}$ can be written as Eq. (13) and Eq. (14). We also use

Online Hard Example Mining (OHEM) [65] to solve the problem of imbalanced foreground and background pixels.

$$\mathcal{L}_{region} = 1 - \frac{2\sum_{i,j}\hat{M}_r(i,j)M_r(i,j)}{\sum_{i,j}\hat{M}_r(i,j)^2 + \sum_{i,j}M_r(i,j)^2}, \tag{13}$$

$$\mathcal{L}_{kernel} = 1 - \frac{2\sum_{i,j}\hat{M}_k(i,j)M_k(i,j)}{\sum_{i,j}\hat{M}_k(i,j)^2 + \sum_{i,j}M_k(i,j)^2}. \tag{14}$$

While calculating $\mathcal{L}_{agg}$, we regard the pixels in M_r but not in M_k as valid pixels, illustrated by M_v in Fig. 4.1. Given N_{tgt} regions of interest in I_{tgt}, the valid pixel in i-th region can be formulated as T_i and kernel as K_i. Then we use Eq. (15) to supervise the overlapped pixels aggregating the proper region, as shown green dashed line of Fig. 4.1.

$$\mathcal{L}_{agg} = \frac{1}{N_{tgt}}\sum_{i=1}^{N_{tgt}}\frac{1}{T_i}\sum_{p\in T_i} ln(1 + D(p, K_i)), \tag{15}$$

$$\mathcal{L}_{dis} = \frac{1}{N_{tgt}(N_{tgt}-1)}\sum_{i=1}^{N_{tgt}}\sum_{j=1,j\neq i}^{N_{tgt}} ln(1 + D(K_i, K_j)), \tag{16}$$

where p is valid pixel of i-th region of interest. $D(\cdot)$ is the distance function. In addition, the cluster centers need to keep discrimination. Therefore, the kernels of different text instances should maintain enough distance as shown red dashline of Fig. 4.1.

5 Experiments

We conduct extensive experiments to analyze the effectiveness of our proposed RoI-Matcher. In this section, we first introduce the implementation details, which are described in Sec. 5.1. Then we compare our RoI-Matcher with other CD-DSS methods due to their similar inputs and outputs. Finally, the ablation studies are shown in Sec. 5.3 to analyze the effectiveness and efficiency of our proposed modules.

5.1 Implementation Details

We implement our RoI-Matcher using the native PyTorch library [67]. Training and testing images are resized to 640×640, with data augmentations including RandomFlip, RandomRotate, and RandomBrightnessContrast. We use AdamW optimizer with a weight decay of 1e-4 and an initial learning rate of 1e-4, adjusted via a step schedule. Experiments run on NVIDIA RTX 2080 Ti GPUs, except for MLLM inference, performed on an NVIDIA RTX A6000.

Table 5.1. Comparison results of RoI-Matcher. † means only samples that fit the output format participate in the calculation during the evaluation.

Methods	Level I		Level II		Level III		Total	
	mIoU (%)	F (%)	mIoU (%)	F (%)	mIoU (%)	F (%)	mIoU (%)	F (%)
Cross-Domain Few-Shot Segmentation								
PATNet [24]	65.1	46.7	60.9	40.1	48.6	15.9	56.8	27.5
RestNet [26]	73.2	63.6	57.8	31.7	48.7	15.1	57.9	29.8
DMTNet [66]	70.3	58.2	61.9	42.7	50.1	20.4	59.5	34.5
ABCD-FSS [27]	40.3	3.8	38.4	16.8	52.0	39.8	43.5	9.2
Multimodal Large Language Model								
Qwen2-VL-7B [28] †	67.4	65.2	63.2	62.3	48.7	33.8	57.3	55.3
InternVL2.5-8B [29] †	67.5	64.8	62.7	60.5	46.0	25.9	53.2	50.5
Class-Agnostic RoI-Matching								
RoI-Matcher (Ours)	92.1	92.2	88.5	84.2	85.0	72.3	87.3	83.9

For comparison experiments, we use CD-FSS and MLLM methods as benchmarks to validate the effectiveness of our RoI-Matcher. For CD-FSS, PATNet [24], RestNet [26], DMTNet [66] and ABCD-FSS [27] all use ResNet50 [68] as the backbone. Because of the difference between task and training dataset, we replace the cross entropy loss of the original model with the Dice loss. For MLLM, we use the open-source method Qwen2-VL-7B [28] and InternVL2.5-8B [29] for comparison. We treat I_{ref}, M_{ref}, and I_{tgt} as inputs, with the goal of predicting M_{tgt} as the output. Moreover, we employ the in-context learning with a reference-target pair to maintain the formatted output.

5.2 Main Results

We use representative methods based on CD-FSS and MLLM to compare the RoI-Matching task performance. Table 5.1 shows the results using mIoU and F-Measure, including different datasets and methods. The CD-FSS method performs well in simple scenes, with ResNet achieving 73.2% mIoU and 63.6% F-Measure. However, its performance declines significantly in challenging scenes, especially at Level III, dropping to 48.7% and 15.1%. While our model also sees some degradation in harder scenarios, it consistently outperforms others across all three levels. As MLLM, Table 5.1 shows Qwen2-VL-7B and InternVL2.5-8B achieves comparable performance in Level I and significantly improves over CD-FSS methods in Level II and Level III, especially in F-Measure. However, apart from performing far below our method in the performance of RoI-Matching, the inference speed of Qwen2-VL-7B(3.121s/img) is much slower than our RoI-Matcher (0.056s/img). By comparing with methods from other tasks, we verify the necessity of RoI Matching and the superior performance of our RoI Matcher. As shown in Fig. 5.1, it achieves strong matching in multi-granular, multi-linguistic and open-set scenarios.

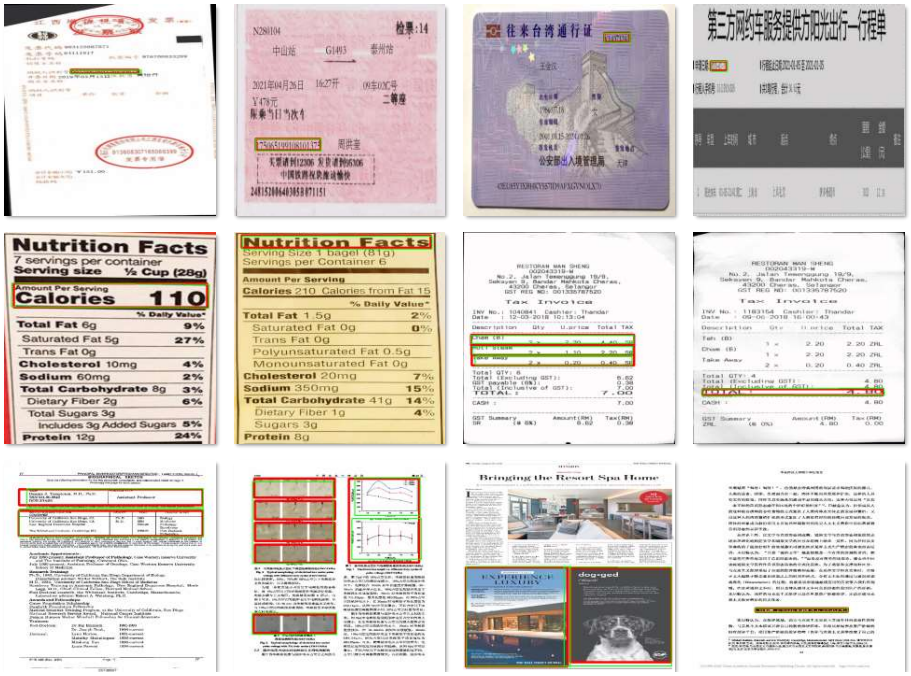


Fig. 5.1. The visualization of RoI-Matcher. The red bounding boxes of each image are ground truths and the green ones are predictions. The difficulty levels I - III are shown from top to bottom in this figure.

Table 5.2. Ablation experiments of Backbones.

Metrics	ResNet-18	ResNet-34	ResNet-50	ResNet-101	Swin-T
mIoU (%)	86.0	85.9	87.3	87.0	78.3
F-Measure (%)	81.0	83.0	83.9	84.2	50.1

5.3 Ablation Studies

Backbones As shown in Table 5.2,We evaluate different ResNet backbones within RoI-Matcher. ResNet-50 outperforms smaller variants like ResNet-18 and ResNet-34, as deeper semantic feature extraction is crucial for this task. However, performance gains from ResNet-50 to ResNet-101 are marginal. We also test Swin-T, which underperforms compared to ResNet, likely due to inferior convergence and adaptability on this task. Therefore, we choose the ResNet-50 as the default setting to balance performance and efficiency.

Training Strategies Due to varying difficulty levels in RoI-Matching-Bench, we compare joint training with curriculum training [69]. Joint training uses all three difficulty subsets simultaneously, while curriculum training progresses from

Table 5.3. Ablation experiments of training strategies.

Training Strategy	mIoU (%)	F-Measure (%)
Joint Training	87.3	83.9
Curriculum Training	79.4	83.2

Table 5.4. Ablation experiments of losses.

Setting	mIoU (%)	F-Measure (%)
CrossEntropy + Dice	85.4	66.8
PANLoss	85.9 (+0.5%)	83.0 (+16.2%)

easy to hard, mimicking human learning. As shown in Table 5.3, although curriculum training yields decent results, joint training performs better in both mIoU and F-Measure. We argue it is because the difficulty we defined in the benchmark is a human semantic difficulty, rather than a natural difficulty like the image from clear to blurry. Therefore, we use the joint training strategy in the following experiments.

Optimization Target In this subsection, we investigate the impact of different loss functions on RoI-Matcher performance, as shown in Table 5.4. Using a combination of cross-entropy and dice loss yields 85.4% mIoU and 66.8% F-Measure. In contrast, applying PANLoss improves results by 0.5% in mIoU and 16.2% in F-Measure. This gain, especially in F-Measure, is attributed to PANLoss's ability to better distinguish overlapping regions.

Efficiency In our baseline, we introduce two efficient modules and conduct ablations to validate their effectiveness, as shown in Table 5.5. The first row reports 4.5M parameters and 149.3 GFlops. Replacing our FPNC with Feature Pyramid Network (FPN) increases parameters by 26.3% and throughput by 55.9%. Removing grid sampling keeps the parameter count unchanged but increases throughput by 114%, significantly reducing inference time. These ablations confirm the efficiency of both design choices.

Pre-mask vs Post-mask We conduct ablation studies to examine the optimal placement of the mask operation relative to our FPNC. As presented in Table 5.6, applying the mask after FPNC (Post-mask) can result in notable information loss in the source domain, indicating that early masking is more effective for preserving semantic features.

6 Limitations

Although we propose a novel task RoI-Matching and design a feasible baseline, several limitations remain. First, our current RoI-Matcher is a 1-shot model,

Table 5.5. Ablation experiments for efficiency.

Settings	Param. (M)	Throughput (GFlops)
RoI-Matcher	34.5	149.3
FPNC → FPN	43.6 (+26.3%)	232.8 (+55.9%)
wo Grid Sampling	34.5 (0%)	319.5 (+114%)

Table 5.6. Comparison between Pre-mask and Post-mask.

Settings	mIoU (%)	F-Measure (%)
Post-mask	87.1	68.4
Pre-mask	87.3 (+0.2%)	83.9 (+22.7%)

unable to self-adjust or refine matching results through multiple visual prompts. Secondly, it is a uni-modal model that does not incorporate any linguistic information. Furthermore, We have not yet attempted repeated inference across multiple RoIs, but we propose a feasible strategy for optimization. When given a multi-channel source mask, RoI-Matcher can generate corresponding masks with the same number of channels for parallel inference. We plan to address these limitations in future work.

7 Conclusion

In this paper, we propose the class-agnostic region-of-interest matching task creatively, to obtain the target corresponding regions from the reference target and customized visual prompt, in a flexible, efficient, multi-granularity, and open-set manner. Then we construct an abundant benchmark RoI-Matching-Bench with three levels of difficulties, which allows researchers to evaluate the performance of RoI-Matching approach under real-world conditions. We also develop an effective and efficient baseline RoI-Matcher. The extensive experiments show that our simple method outperforms other CD-FSS and MLLM methods. Several ablations also indicate our subtle designs are effective and efficient. In the future, we hope to enhance our RoI-Matcher in the perspective of multi-modal and multi-interactive.

Acknowledgments

Supported by the National Natural Science Foundation of China (Grant NO 62376266 and 62406318), Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing, China.

References

1. Yan Shu, Weichao Zeng, Fangmin Zhao, Zeyu Chen, Zhenhang Li, Xiaomeng Yang, Yu Zhou, Paolo Rota, Xiang Bai, Lianwen Jin, Xu-Cheng Yin, and Nicu Sebe. Visual text processing: A comprehensive review and unified evaluation, 2025.

2. Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. TextCtrl: Diffusion-based scene text editing with prior guidance control. In *NeurIPS*, pages 138569–138594, 2024.
3. Tianjiao Cao, Jiahao Lyu, Weichao Zeng, Weimin Mu, and Yu Zhou. The devil is in fine-tuning and long-tailed problems: A new benchmark for scene text detection. In *IJCAI*, 2025.
4. Yan Shu, Wei Wang, Yu Zhou, Shaohui Liu, Aoting Zhang, Dongbao Yang, and Weipinng Wang. Perceiving ambiguity and semantics without recognition: An efficient and effective ambiguous scene text detector. In *ACM MM*, page 1851–1862, 2023.
5. Yudi Chen, Wei Wang, Yu Zhou, Fei Yang, Dongbao Yang, and Weiping Wang. Self-training for domain adaptive scene text detection. In *ICPR*, pages 850–857. IEEE, 2021.
6. Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. SEED: Semantics enhanced encoder-decoder framework for scene text recognition. In *CVPR*, pages 13528–13537, 2020.
7. Zhi Qiao, Yu Zhou, Jin Wei, Wei Wang, Yuan Zhang, Ning Jiang, Hongbin Wang, and Weiping Wang. PIMNet: A parallel, iterative and mimicking network for scene text recognition. In *ACM MM*, pages 2046–2055, 2021.
8. Xiaomeng Yang, Zhi Qiao, and Yu Zhou. IPAD: Iterative, parallel, and diffusion-based network for scene text recognition. *IJCV*, 2025.
9. Yifei Zhang, Chang Liu, Jin Wei, Xiaomeng Yang, Yu Zhou, Can Ma, and Xiangyang Ji. Linguistics-aware masked image modeling for self-supervised scene text recognition. In *CVPR*, pages 9318–9328, 2025.
10. Huawen Shen, Xiang Gao, Jin Wei, Liang Qiao, Yu Zhou, Qiang Li, and Zhanzhan Cheng. Divide rows and conquer cells: Towards structure recognition for large tables. In *IJCAI*, pages 1369–1377, 2023.
11. Jiahao Lyu, Wei Wang, Dongbao Yang, Jinwen Zhong, and Yu Zhou. Arbitrary reading order scene text spotter with local semantics guidance. In *AAAI*, volume 39, pages 5919–5927, 2025.
12. Wei Wang, Yu Zhou, Jiahao Lv, Dayan Wu, Guoqing Zhao, Ning Jiang, and Weipinng Wang. TPSNet: Reverse thinking of thin plate splines for arbitrary shape scene text representation. In *ACM MM*, pages 5014–5025, 2022.
13. Jiahao Lyu, Jin Wei, Gangyan Zeng, Zeng Li, Enze Xie, Wei Wang, Can Ma, and Yu Zhou. TextBlockV2: Towards precise-detection-free scene text spotting with pre-trained language model. *TOMM*, 2025.
14. Gengluo Li, Huawen Shen, and Yu Zhou. Beyond cropped regions: New benchmark and corresponding baseline for chinese scene text retrieval in diverse layouts. In *ICML*, 2025.
15. Gangyan Zeng, Yuan Zhang, Jin Wei, Dongbao Yang, Peng Zhang, Yiwen Gao, Xugong Qin, and Yu Zhou. Focus, distinguish, and prompt: Unleashing clip for efficient and flexible scene text retrieval. In *ACM MM*, pages 2525–2534, 2024.
16. Huawen Shen, Gengluo Li, Jinwen Zhong, and Yu Zhou. LDP: Generalizing to multilingual visual information extraction by language decoupled pretraining. In *AAAI*, volume 39, pages 6805–6813, 2025.
17. Gangyan Zeng, Yuan Zhang, Yu Zhou, and Xiaomeng Yang. Beyond OCR+ VQA: Involving ocr into the flow for robust and accurate textvqa. In *ACM MM*, pages 376–385, 2021.
18. Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *ICDAR*, pages 1015–1022. IEEE, 2019.

19. Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *CVPR*, pages 5315–5324, 2017.
20. Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. VSR: a unified framework for document layout analysis combining vision, semantics and relations. In *ICDAR*, pages 115–130. Springer, 2021.
21. Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, and Xing Sun. Attention where it matters: Rethinking visual document understanding with selective region concentration. In *ICCV*, pages 19517–19527, 2023.
22. Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *AAAI*, volume 36, pages 10767–10775, 2022.
23. Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. StrucText: Structured text understanding with multi-modal transformers. In *ACM MM*, pages 1912–1920, 2021.
24. Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *ECCV*, pages 73–90. Springer, 2022.
25. Antonio Tavera, Fabio Cermelli, Carlo Masone, and Barbara Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. In *WACV*, pages 1626–1635, 2022.
26. Xinyang Huang, Chuang Zhu, and Wenkai Chen. Restnet: Boosting cross-domain few-shot segmentation with residual transformation network. In *BMVC*, 2023.
27. Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *CVPR*, pages 23605–23615, 2024.
28. Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv*, 2024.
29. Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024.
30. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016.
31. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017.
32. Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *CVPR*, pages 15138–15147, 2023.
33. Ayan Banerjee, Sanket Biswas, Josep Lladós, and Umapada Pal. Swindocsegmenter: An end-to-end unified domain adaptive transformer for document instance segmentation. In *ICDAR*, pages 307–325. Springer, 2023.
34. Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023.

35. Subhajit Maity, Sanket Biswas, Siladittya Manna, Ayan Banerjee, Josep Lladós, Saumik Bhattacharya, and Umapada Pal. Selfdocseg: A self-supervised vision-based approach towards document segmentation. In *ICDAR*, pages 342–360. Springer, 2023.
36. Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020.
37. Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *ACM SIGKDD*, pages 1192–1200, 2020.
38. Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*, pages 2579–2591, 2021.
39. Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for document ai with unified text and image masking. In *ACM MM*, pages 4083–4091, 2022.
40. Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: Bert pre-training of image transformers. In *ICLR*, 2022.
41. Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. In *ACM MM*, pages 3530–3539, 2022.
42. Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pretraining framework for document understanding. *NeurIPS*, 34:39–50, 2021.
43. Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. StrucTexTv2: Masked visual-textual prediction for document image pre-training. In *ICLR*, 2023.
44. Weizhao He, Yang Zhang, Wei Zhuo, Linlin Shen, Jiaqi Yang, Songhe Deng, and Liang Sun. Apseg: auto-prompt network for cross-domain few-shot semantic segmentation. In *CVPR*, pages 23762–23772, 2024.
45. Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. Layoutreader: Pre-training of text and layout for reading order detection. *arXiv preprint arXiv:2108.11591*, 2021.
46. Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *EMNLP*, 2023.
47. Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021.
48. Haoyu Cao, Xin Li, Jiefeng Ma, Deqiang Jiang, Antai Guo, Yiqing Hu, Hao Liu, Yinsong Liu, and Bo Ren. Query-driven generative network for document information extraction in the wild. In *ACM MM*, pages 4261–4271, 2022.
49. Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *CVPR*, pages 19254–19264, 2023.
50. Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, pages 18893–18912. PMLR, 2023.

51. Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, and Radu Soricut. Prestu: Pre-training for scene-text understanding. In *ICCV*, pages 15270–15280, 2023.
52. Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, pages 498–517. Springer, 2022.
53. Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *CVPR*, pages 15641–15653, 2024.
54. Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. Icdar 2023 competition on structured text extraction from visually-rich document images. In *ICDAR*, pages 536–552. Springer, 2023.
55. Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *ICDAR*, pages 36–53. Springer, 2023.
56. Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. Towards robust visual information extraction in real world: new dataset and novel solution. In *AAAI*, volume 35, pages 2738–2745, 2021.
57. Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis. In *ICCV*, pages 19462–19472, 2023.
58. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
59. Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, volume 34, pages 11474–11481, 2020.
60. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.
61. Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *ICCV*, pages 8440–8449, 2019.
62. Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *CVPR*, pages 9336–9345, 2019.
63. Bala R Vatti. A generic solution to polygon clipping. *Communications of the ACM*, 35(7):56–63, 1992.
64. Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571. Ieee, 2016.
65. Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.
66. Minh-Thien Duong, Seongsoo Lee, and Min-Cheol Hong. Dmt-net: Deep multiple networks for low-light image enhancement based on retinex model. *IEEE Access*, 11:132147–132161, 2023.
67. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al.

Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.
68. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
69. Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE TPAMI*, 44(9):4555–4576, 2021.

2027届本科生公能素质评价加分材料

	类别	分数（自填）	审核结果
思想成长与身心发展	理想信念坚定	5	
	热心服务集体	0	
	先锋作用突出	0	
志愿服务与社会实践	志愿服务	3	
	社会实践	0.75	
艺体素质与技能特长	代表学院参与文体竞赛	0	
	文体竞赛获奖	0	
到国际组织挂职实习	---	0	
总计	---	8.75	

材料统计截止时间：

2021级 2024年6月30日 ☐
2022级 2025年6月29日 ☐
2023级 2026年7月05日 ☐

学号： 2312141

姓名： 张德民

所在班级： 信息安全班

（一）思想成长与身心发展（上限10分）

1、理想信念坚定（上限5分）

在校期间思想表现情况自述：

在校期间，我主动加强思想政治理论学习，认真学习党的创新理论与各项方针政策，树立正确的世界观、人生观和价值观，政治立场坚定，思想觉悟较高。我日常生活中尊敬师长，待人真诚友善，能够团结同学，主动配合集体工作，恪守诚信原则，乐于帮助身边同学解决学习生活中的问题。我自觉遵守国家法律法规以及学校各项管理规定，严于律己，品行端正，在校期间无任何违纪违规行为，未受到过任何处分。

在专业学习上，我态度端正、刻苦钻研，扎实掌握计算机与信息安全相关专业理论与实践技能。我注重提升自身科研素养，积极参与科研相关学习实践，对待科研严谨踏实，具备较好的探索钻研精神，对科研工作抱有浓厚兴趣，并取得了一定的成果，我希望未来能够持续深耕专业领域，努力实现自身价值。

本人签字：

日期：

辅导员评分：

辅导员签字：

日期：

有无违反法律法规： 无

有无违反学校纪律处分规定： 无

有无发生思想品德问题并造成不良影响： 无

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）志愿服务时长证明

基本资料

学号：	2312141
姓名：	张德民
身份证号：	410882200411240174
志愿者编号：	410882001109800972
出生年月：	2004-11-24
学院：	密码与网络空间安全学院
专业：	信息安全
年级：	2023级
入学时间：	2023-09-01
毕业时间：	2027-07-01
政治面貌：	共青团员
手机号：	18339765000
服务时长：	81.5 小时
校外服务时长：	0 小时

编辑

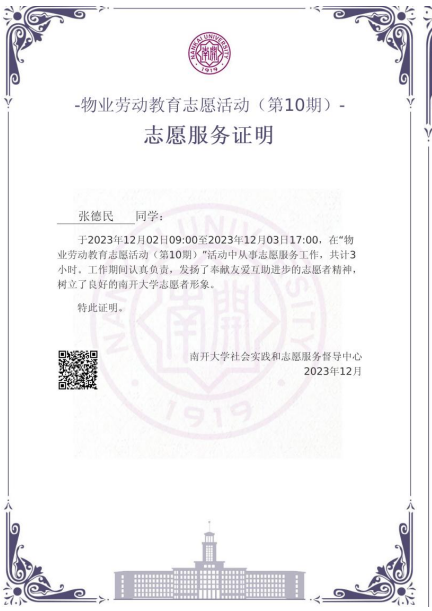
前三学年内志愿服务总时长（小时）： 81.5

除公能实践课要求外志愿服务总时长（小时）： 68

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）除公能实践课课要求18学时外，剩余志愿服务时长证明（需在统计时间范围内且为志愿服务平台导出证明）



-物业劳动教育志愿活动（第10期）-

志愿服务证明

张德民 同学：

于2023年12月02日09:00至2023年12月03日17:00，在“物业劳动教育志愿活动（第10期）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2023年12月

志愿服务时长（小时）：

3

-“智慧助老，温暖相伴”——第一期-

志愿服务证明

张德民 同学：

于2024年10月19日09:00至2024年10月19日11:00，在““智慧助老，温暖相伴”——第一期”活动中从事志愿服务工作，共计2小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2024年10月

志愿服务时长（小时）：

2

-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2024年11月09日09:00至2024年11月10日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2024年11月

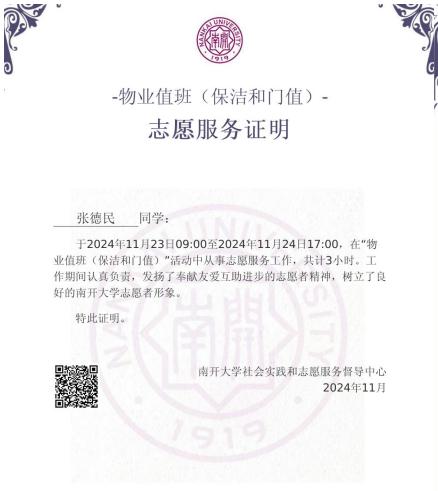
志愿服务时长（小时）：

3

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）除公能实践课课要求18学时外，剩余志愿服务时长证明（需在统计时间范围内且为志愿服务平台导出证明）



-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2024年11月23日09:00至2024年11月24日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2024年11月

志愿服务时长（小时）：

3

-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2024年11月30日09:00至2024年12月01日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2024年12月

志愿服务时长（小时）：

3

-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2025年03月22日09:00至2025年03月23日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2025年3月

志愿服务时长（小时）：

3

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）除公能实践课课要求18学时外，剩余志愿服务时长证明（需在统计时间范围内且为志愿服务平台导出证明）

-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2025年09月27日09:00至2025年09月27日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计2.5小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

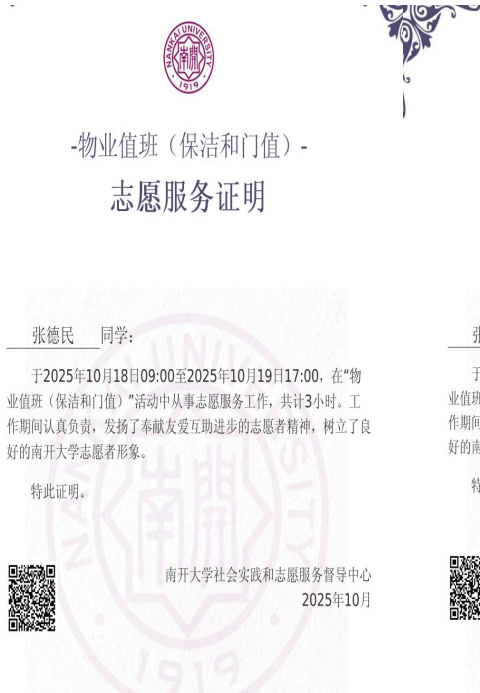
特此证明。

南开大学社会实践和志愿服务督导中心

2025年9月

志愿服务时长（小时）：

2.5



-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2025年10月18日09:00至2025年10月19日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

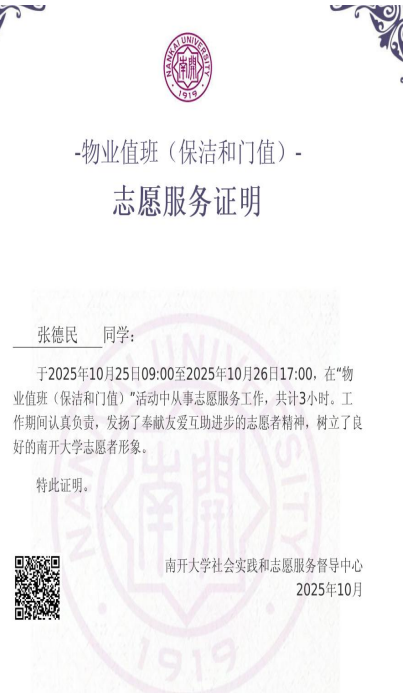
特此证明。

南开大学社会实践和志愿服务督导中心

2025年10月

志愿服务时长（小时）：

3



-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2025年10月25日09:00至2025年10月26日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2025年10月

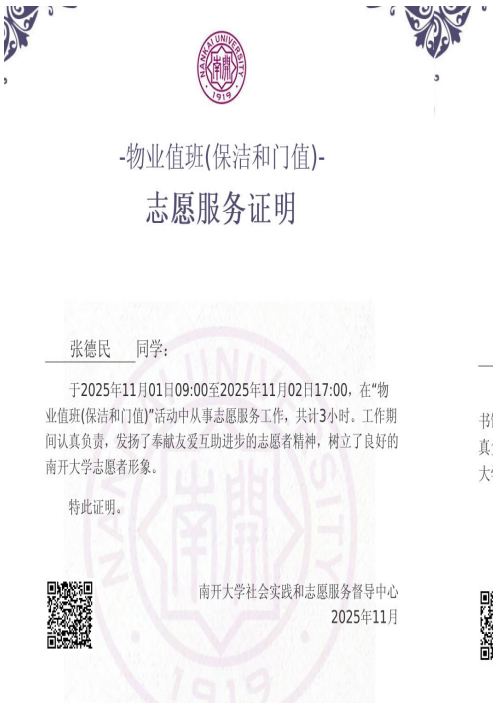
志愿服务时长（小时）：

3

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）除公能实践课课要求18学时外，剩余志愿服务时长证明（需在统计时间范围内且为志愿服务平台导出证明）



-物业值班(保洁和门值)-

志愿服务证明

张德民 同学：

于2025年11月01日09:00至2025年11月02日17:00，在“物业值班(保洁和门值)”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

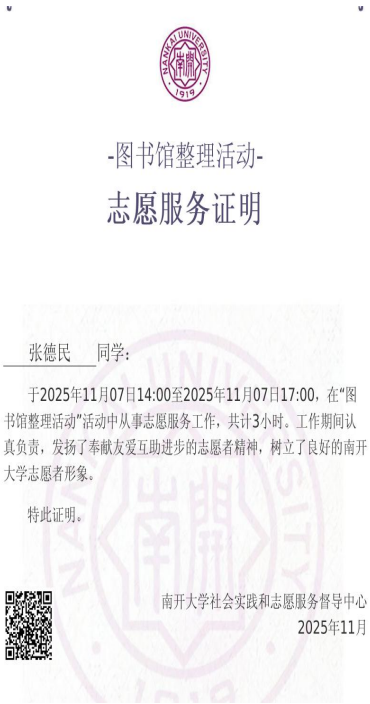
特此证明。

南开大学社会实践和志愿服务督导中心

2025年11月

志愿服务时长（小时）：

3



-图书馆整理活动-

志愿服务证明

张德民 同学：

于2025年11月07日14:00至2025年11月07日17:00，在“图书馆整理活动”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

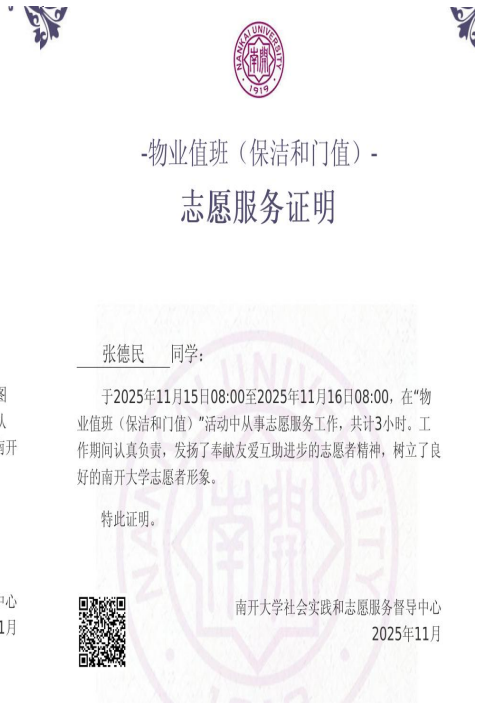
特此证明。

南开大学社会实践和志愿服务督导中心

2025年11月

志愿服务时长（小时）：

3



-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2025年11月15日08:00至2025年11月16日08:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2025年11月

志愿服务时长（小时）：

3

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）除公能实践课课要求18学时外，剩余志愿服务时长证明（需在统计时间范围内且为志愿服务平台导出证明）

-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2025年11月29日09:00至2025年11月30日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2025年11月

志愿服务时长（小时）：

3

-图书馆整理活动-

志愿服务证明

张德民 同学：

于2026年03月12日14:00至2026年03月12日17:00，在“图书馆整理活动”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

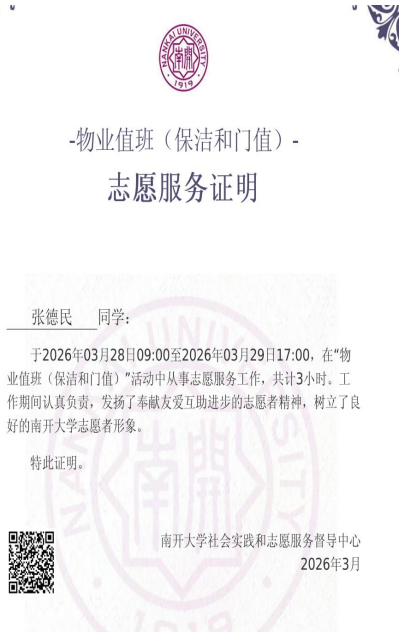
特此证明。

南开大学社会实践和志愿服务督导中心

2026年3月

志愿服务时长（小时）：

3



-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2026年03月28日09:00至2026年03月29日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2026年3月

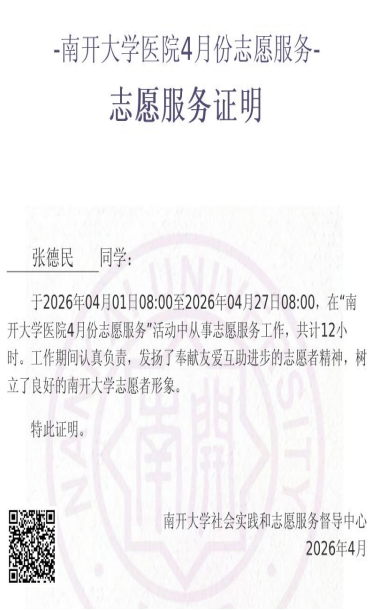
志愿服务时长（小时）：

3

（二）志愿服务与社会实践（上限6分）

1、志愿服务

（1）除公能实践课课要求18学时外，剩余志愿服务时长证明（需在统计时间范围内且为志愿服务平台导出证明）



-南开大学医院4月份志愿服务-

志愿服务证明

张德民 同学：

于2026年04月01日08:00至2026年04月27日08:00，在“南开大学医院4月份志愿服务”活动中从事志愿服务工作，共计12小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

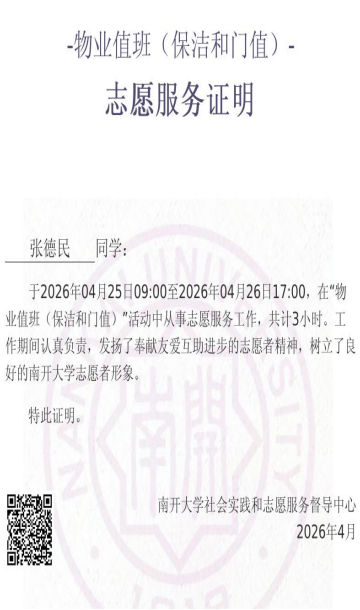
特此证明。

南开大学社会实践和志愿服务督导中心

2026年4月

志愿服务时长（小时）：

12



-物业值班（保洁和门值）-

志愿服务证明

张德民 同学：

于2026年04月25日09:00至2026年04月26日17:00，在“物业值班（保洁和门值）”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

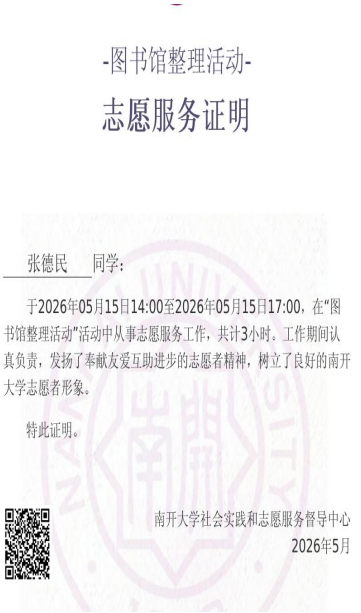
特此证明。

南开大学社会实践和志愿服务督导中心

2026年4月

志愿服务时长（小时）：

3



-图书馆整理活动-

志愿服务证明

张德民 同学：

于2026年05月15日14:00至2026年05月15日17:00，在“图书馆整理活动”活动中从事志愿服务工作，共计3小时。工作期间认真负责，发扬了奉献友爱互助进步的志愿者精神，树立了良好的南开大学志愿者形象。

特此证明。

南开大学社会实践和志愿服务督导中心

2026年5月

志愿服务时长（小时）：

3

（二）志愿服务与社会实践（上限6分）

2、社会实践（上限为2分）

（1）社会实践次数证明

活动名称	活动时间	认定时长	发布人	操作
赓续红色根脉——中国共产党历史展览馆与党的百年奋斗史寻访实践	2026-06-27 08:00到 2026-06-27 16:00	7 小时	杜泽琦	查看 打印
探索蓝色星球，共筑海洋强国梦	2025-03-08 11:00到 2025-03-08 17:00	6 小时	李听泉	查看
2025年“梦圆南开 心系母校”的团体实践	2024-12-01 08:00到 2025-02-28 08:00	22 小时	张颖	查看 打印
2024年“梦圆南开 心系母校”的团体实践	2023-12-01 08:00到 2024-02-28 08:00	20 小时	张颖	查看 打印

大一寒假是否参加社会实践：是

大二寒假是否参加社会实践：是

大三寒假是否参加社会实践：否

大一暑假是否参加社会实践：否

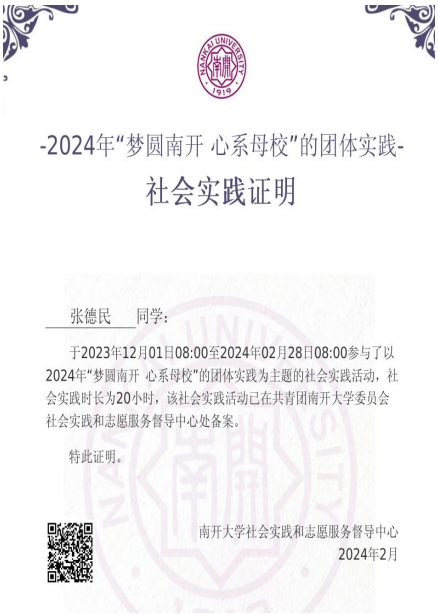
大二暑假是否参加社会实践：否

大三暑假是否参加社会实践：是

（二）志愿服务与社会实践（上限6分）

2、社会实践（上限为2分）

（1）社会实践证明



-2024年“梦圆南开 心系母校”的团体实践-

社会实践证明

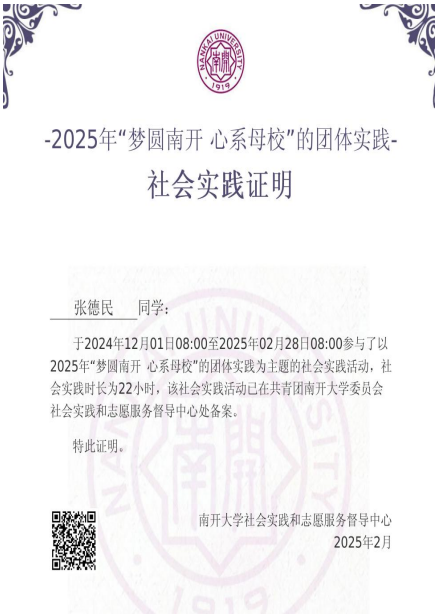
张德民 同学：

于2023年12月01日08:00至2024年02月28日08:00参与了以2024年“梦圆南开 心系母校”的团体实践为主题的社会实践活动，社会实践时长为20小时，该社会实践活动已在共青团南开大学委员会社会实践和志愿服务督导中心处备案。

特此证明。

南开大学社会实践和志愿服务督导中心

2024年2月



-2025年“梦圆南开 心系母校”的团体实践-

社会实践证明

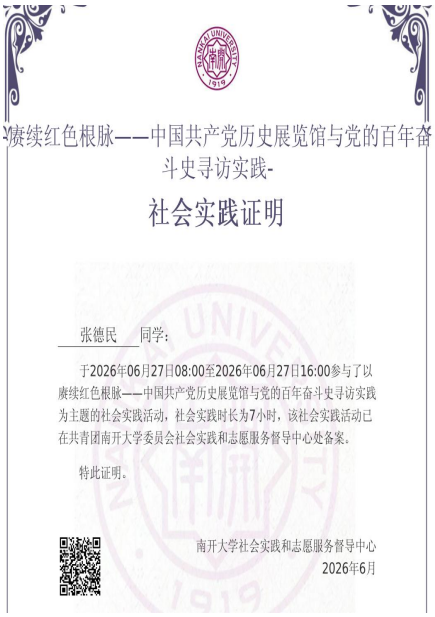
张德民 同学：

于2024年12月01日08:00至2025年02月28日08:00参与了以2025年“梦圆南开 心系母校”的团体实践为主题的社会实践活动，社会实践时长为22小时，该社会实践活动已在共青团南开大学委员会社会实践和志愿服务督导中心处备案。

特此证明。

南开大学社会实践和志愿服务督导中心

2025年2月



-赓续红色根脉——中国共产党历史展览馆与党的百年奋斗史寻访实践-

社会实践证明

张德民 同学：

于2026年06月27日08:00至2026年06月27日16:00参与了以赓续红色根脉——中国共产党历史展览馆与党的百年奋斗史寻访实践为主题的社会实践活动，社会实践时长为7小时，该社会实践活动已在共青团南开大学委员会社会实践和志愿服务督导中心处备案。

特此证明。

南开大学社会实践和志愿服务督导中心

2026年6月