SRE: A Class-missing Semi-supervised Document Key Information Extraction via Synergistic Refinement Estimation

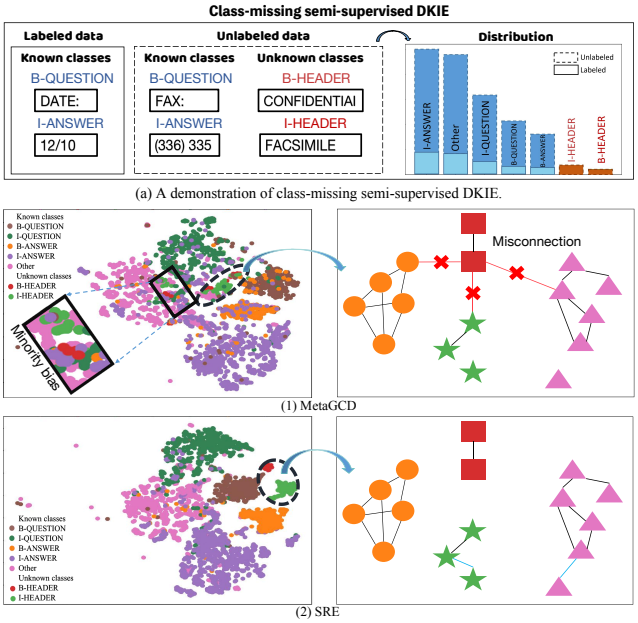
Pengcheng Guo, Yonghong Song, *Member, IEEE,* Boyu Wang, *Member, IEEE,* Yankai Cao, Jiayang Ren, Chaojie Ji, Jiahao Liu, Qi Zhang, Qiangqiang Mao

***Abstract*—Current methods for document key information extraction (DKIE) rely heavily on labeled data with high annotation costs. To mitigate this issue, semi-supervised learning (SSL) paradigm that utilizes unlabeled document samples, has gained broad attention in DKIE. However, existing SSL methods require labeled and unlabeled data to share identical label space, which is impractical in many DKIE tasks (i.e., some unlabeled samples do not belong to any known classes in the labeled set). In this paper, we formulate this problem as class-missing semi-supervised (CMSS) DKIE. In DKIE, unknown classes usually belong to minority and fine-grained categories, intensifying the misconnections between known and unknown classes and making CMSS more challenging. To address this issue, we propose Synergistic Refinement Estimation (SRE), a progressive prototype estimation scheme that alleviates the unknown classes bias to the majority known classes on long-tailed unlabeled data. Furthermore, dynamic threshold hash rectification and structural calibration mechanisms are proposed to correct connections between fine-grained classes. Extensive experimental results demonstrate that SRE surpasses existing state-of-the-art methods on two DKIE benchmarks. Code is available at https://github.com/anonymoulink/SRE_DKIE.**

***Index Terms*—Information extraction, Class-missing semi-supervised, Calibration, Fine-grained classes.**

I. Introduction

DOCUMENT Key Information Extraction (DKIE) aims to extract vital information from digital documents, which are widely used in many industries [1], [2]. Different from general image classification datasets, DKIE datasets contain not only visual information but also additional information, such as contextual content, layout and structural features [3], [4]. Several deep learning methods have achieved remarkable success in the DKIE task by incorporating multi-source information [3], [5]. However, they all rely on enormous labeled training samples to achieve state-of-the-art performance, which is expensive and hinders the widespread application of existing deep learning-based DKIE methods.



(a) A demonstration of class-missing semi-supervised DKIE.

(b) The t-SNE visualization of features and corresponding graph structure for different methods on FUNSD.

Fig. 1. (a) A demonstration of class-missing semi-supervised DKIE, where known classes belong to majority classes and unknown classes belong to minority classes. (b) t-SNE visualization of MetaGCN and SRE on FUNSD in latent representations (left). The representations represented by the dashed circles in (1) reflect misconnections within fine-grained unknown categories on the graph (right). The square dashed box in (1) reflects the bias of the minority unknown categories toward the head known category on the unlabeled data.

Semi-supervised learning (SSL) methods [6], [7] offer an effective approach to alleviate the burden of labeling by leveraging a large corpus of unlabeled samples. However, conventional SSL approaches typically assume that both labeled and unlabeled data share the identical label space [8], which is impractical for DKIE, since unlabeled DKIE data inevitably includes many classes unknown to the labeled ones. For instance, in the DKIE task of receipt classification, discount information is relatively rare and never exists in the labeled receipts, and therefore belongs to an unknown class in the unlabeled data.

In this paper, we formulate such a problem as class-missing SSL (CMSS) DKIE, assuming that unlabeled data contains both known and unknown classes. Additionally, in DKIE, the unknown classes usually belong to the minority and fine-grained classes, which makes CMSS even more challenging. **1.** For example, as shown in Fig. 1-a, the unlabeled receipt

Pengcheng Guo, Yonghong Song, Jiahao Liu and Qi Zhang are with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China. (Email: bedlexmunaxl@stu.xjtu.edu.cn; songyh@mail.xjtu.edu.cn; para15291067561@163.com

Boyu Wang is with the Department of Computer Science and the Brain Mind Institute, University of Western Ontario, London, ON N6A 3K7, Canada, and also with the Vector Institute, Toronto, ON M5G 1M1, Canada. E-mail: bwang@csd.uwo.ca

Yankai Cao, Jiayang Ren and Qiangqiang Mao are with the Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC V6T 1Z3, Canada (Email: yankai.cao@ubc.ca, rjy12307@mail.ubc.ca, maoq@mail.ubc.ca). Chaojie Ji is with the Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada (Email: chaojiej@math.ubc.ca)

contains two unknown and *minority* classes "I-HEADER" and "B-HEADER", which can be dominated by majority classes (e.g., "I-ANSWER"). Note that this problem is fundamentally different from the conventional setting of imbalanced classification [9], [10], where the minority classes are known (i.e., "I-HEADER" and "B-HEADER" are labeled). **2.** Moreover, structured annotation in the DKIE data leads to *fine-grained yet indiscernible* classes. For example, "B-HEADER" and "I-HEADER" denote the boundary and inside of the "HEADER" entity, and their semantic information is close to each other. Consequently, existing methods (e.g., MetaGCD [11]) will produce erroneous connections between known and unknown classes on DKIE data since they ignore the underly semantic information of DKIE, as shown in Fig. 1-b-(1).

Existing works [9], [10], [12] have explored the long-tailed SSL setting, but they are not applicable to CMSS as they assume that both labeled and unlabeled data share the same class space. While some recent works [11], [13], [14] explore the connections between known and unknown classes by representation learning, they primarily focus on visual feature differences rather than semantic or structural information of DKIE. Consequently, they cannot work well on DKIE with fine-grained classes.

To address the aforementioned issues, we propose a **S**ynergistic **R**efinement **E**stimation (SRE) method for CMSS DKIE, which has the following desirable properties:

Alleviating the bias towards the majority known classes. We propose the **P**rogressive **P**rototype **E**stimation (PPE) scheme to alleviate the bias of minority unknown classes towards majority known classes. Specifically, a prototype (i.e., class center) alignment method is designed to estimate the frequency of each class on unlabeled data, and then an adaptive factor is introduced to encourage the model to produce a more balanced probability distribution of known and unknown classes. Furthermore, a post-processing scheme is designed to enhance the confidence of pseudo-labels by using posterior probability for minority classes.

Calibrating connections between fine-grained classes. We introduce a **S**ynergistic **R**efinement (SR) mechanism to calibrate connections between fine-grained classes. More specifically, dynamic threshold hash rectification and structural calibration mechanisms are proposed to synergistically correct connections between fine-grained classes.

Extensive comparisons on two benchmarks show that SRE significantly outperforms others in the CMSS DKIE setting.

II. Related Work

A. Document Key Information Extraction

Recently, DKIE has attracted a wide range of attention because it can reduce labor costs for companies dealing with large business files, such as purchase orders, business insurances, shopping bills, etc. Fig. 8 describes the pipeline for DKIE tasks in Appendix A1. DKIE models mainly contain three categories: Grid-based approaches [15], [16], Graph Neural Network (GNN)-based approaches [17]–[19] and transformer-based pre-training approaches [5], [20], [21]. Grid-based approaches [15], [16] mainly exploit 2D document representation, where texts are categorized into different entity types by being encoded into segment embeddings. GNN-based approaches [17]–[19] employ text representation as nodes and the GNN's edges relationships to analyze the connections between the various entities. Transformer-based pre-training approaches [5], [20], [21], utilizing text, layout, and image information, has significantly propelled DKIE research forward, making these methods a viable solution for diverse DKIE applications. Among these, transformer-based pre-training approaches [5], [20], [21] have demonstrated their efficacy across a range of DKIE challenges. For example, LayoutLM [4], a BERT-like transformer model, was pioneering in integrating 2D spatial embeddings with each token. Subsequently, LayoutLMv2 [21] enhanced the original model by incorporating both visual and textual information during the pre-training phase. LayoutLMv3 [5] is a widely-used document pre-trained model, which utilizes masked language modeling, masked image modeling, and word-patch alignment to learn multi-modal representations. While significant progress, current models are constrained by the insufficient availability of high-quality annotations. To overcome this limitation, we introduces a new approach for class-misssing semi-supervised DKIE, which leverages synergistic refinement estimation to identify known and unknown classes on unlabeled dataset.

B. Semi-supervised Learning

SSL aims to combine limited labeled data with massive unlabeled samples to reduce the cost of annotation [22]–[24]. Existing strategies for SSL fall into three categories, pseudo-labeling [25]–[27], consistency regularization [28], [29], co-training [29], etc. Recently, SSL methods based on pseudo-labeling have achieved great progress, using the predictions of the model to explicitly produce a pseudo-label for unlabeled samples. However, most existing SSL algorithms assume the datasets are class-balanced, which holds untrue in realistic scenarios. Recently, long-tailed SSL methods [9], [10], [12], [30] have received significant attention for its practicality in many real-world tasks. For example, DARP [9] and CReST [10] eliminate biased pseudo-labels by distribution alignment. ACR [30] achieves adaptive refinement of pseudo-labels for various distributions through a adaptive consistency regularizer by estimating the true class distribution. However, these efforts assume that both labeled and unlabeled data are from known class distribution, resulting in performance degradation in CMSS DKIE task. In contrast, our method designs the SR mechanism to calibrate misconnections for both known and unknown classes. For further details about these methods and differences between CMSS and SSL problems, please refer to the Appendix A2.

C. Novel Class Discovery (NCD)

NCD assumes that unlabeled set and labeled set do not have any class overlap, has gained attention [31]–[33]. UNO [32] unifies multiple loss functions to enhance the information communication between known classes and unknown classes. RankStats [31] utilizes ranking statistics to obtain pseudo-labels for the classification heads. While recent works [11],

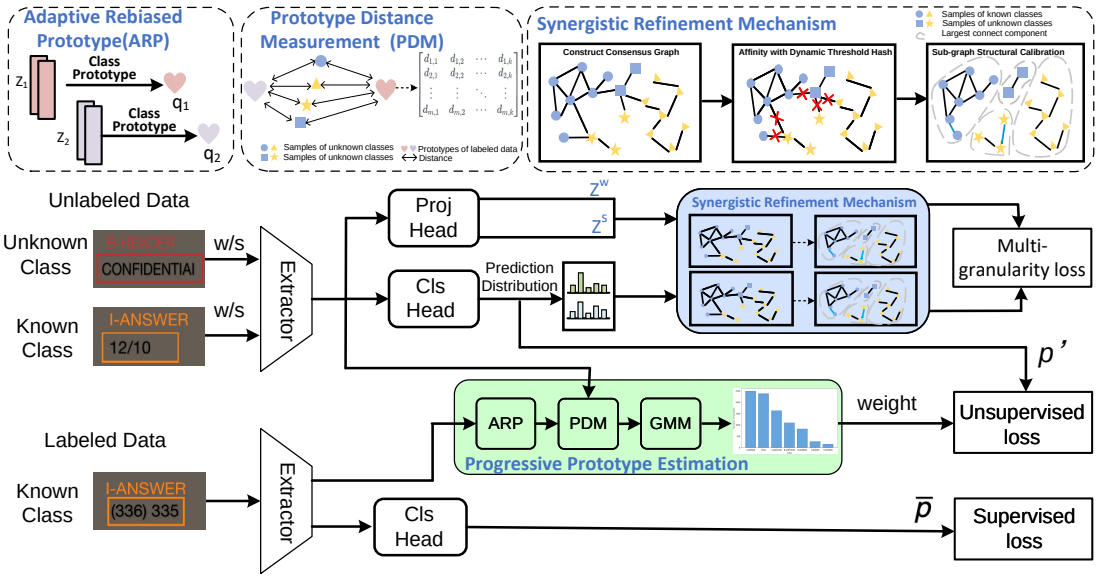


Fig. 2. Framework of the proposed SRE with progressive prototype estimation (PPE) and synergistic refinement (SR) mechanism. $\bar{p}$ is the prediction on labeled set and p' denotes prediction on unlabeled samples. Cls is a classification head and $Proj$ is the projection head. Multi-grnularity loss denotes a Multi-granularity contrastive learning loss. w and s denote weak and strong augmentation, respectively.

[13], [14] relax the assumption in NCD and can address unknown classes on unlabeled data, they fail to calibrate incorrect connections of unknown classes belonging to minority and fine-grained classes. Generalized Category Discovery (GCD) [14] extends NCD, by addressing the classification of images in a dataset where some have known labels, aiming to discover novel categories in unlabeled datasets using knowledge from labeled images. Open-world semi-supervised learning (ORCA) [13] considers the problem from a semi-supervised learning perspective and introduces an adaptive margin loss for better intra-class separability for both known and unknown classes. MetaGCD [11] proposes a meta-learning-based optimization strategy for incremental learning with reduced forgetting, along with soft neighborhood contrastive learning to adaptively support instances using their neighboring data. Different from our setting, they fail to discover unknown classes on unlabeled entities or calibrate misconnection among semantic or structural information of unknown classes belong to minority and fine-grained classes.

D. Open-Set Semi-Supervised Learning

Open-set SSL relaxes the assumption of SSL and considers a more practical scenario that training data could contain unkonwn class unlabeled examples. This task assumes test examples are from known classes, focusing on minimizing unknown class impact to maintain robustness in known classes. Many open-set SSL methods have been proposed in recent years [34]–[36], such as T2T [35] , OpenMatch [36], SAFESTUDENT [37], and IOMatch [38]. T2T [35] presents a cross-modal matching mechanism via binary detectors to detect unknown classes. OpenMatch [36], which uses a group of one-vs-all classifiers as the outlier detector. SAFESTUDENT [37] introduces an energy-discrepancy score as a substitute for confidence. IOMatch [38] introduces a novel unified paradigm for jointly utilizing open-set unlabeled data without explicitly distinguishing between inliers and outliers. Different from open-set SSL, which aims to detect test set elements as not belonging to known classes without further classification, CMSS needs to classify these unknown classes. For more details about these methods and the differences between CMSS and SSL problems, please refer to Appendix C.

III. Method

A. Problem Description

In CMSS DKIE, the training dataset includes the limited labeled data set $D_l = \{(x_i^l, y_i^l)\}_{i=1}^n$ consisting of n samples with labels belonging to C^l classes, C^l is the set of known classes. An unlabeled data set $D_u = \{x_i^u\}_{i=1}^m$ consisting m unlabeled samples, each of which belongs to one of the classes in C^u. C^u is the set of classes in unlabeled training set. Generally, $m \gg n$. Here, $C^l \subset C^u$, then the set of unknown classes denoted by $C^n = C^u / C^l$. The classes included in the testing set are denoted as C^u.

B. Method Overview

The overall framework is shown in Fig. 2, which has two main components: 1) Progressive Prototype Estimation (PPE) module, which is introduced to alleviate over-biasing the predictions on unknown samples towards majority known classes. 2) Synergistic Refinement (SR) mechanism, which is introduced to calibrate connections among fine-grained classes. SR mechanism includes three consecutive stages: constructing consensus graph, updating affinity with Dynamic Threshold Hash, and creating sub-graph with structural calibration.

The training procedure is shown in Alg. 1. First, given a sample x, a shared DKIE extractor $\mathcal{F}(\cdot)$ is used to extract the feature z. Then, a classification head $\Phi_{cls}(\cdot)$ generates the class probability $p_{cls} = \Phi_{cls}(z)$. Meanwhile, PPE module (in Section III-C) is utilized to construct prototypes on labeled instances and estimate known and unknown classes on unlabeled instances. Subsequently, for an unlabeled instance

x^u, we utilize a multi-layer perceptron (MLP) projection head $\Phi_{pro}(\cdot)$ to obtain weakly augmented and strongly augmented representations, denoted as $z^w = \Phi_{pro}(\mathcal{F}(\mathcal{A}_w(x^u)))$, $z^s = \Phi_{pro}(\mathcal{F}(\mathcal{A}_s(x^u)))$ which are used for SR mechanism (in Section III-D). Here, SRE employs both weak augmentation $\mathcal{A}_w(\cdot)$ and strong augmentations $\mathcal{A}_s(\cdot)$.

C. Class Estimation via Progressive Prototype Estimation

To mitigate over-biasing the predictions on unknown samples towards majority known classes, we propose a progressive prototype estimation module. Firstly, we estimate the frequency of each class on unlabeled data with prototype alignment way. Subsequently, incorporating the adaptive factor into the consistency loss based on such estimation. Moreover, a post-processing scheme is introduced to further improve the confidence of minority classes. In our work, the prototypes can be interpreted as the class centers for each of the categories.

Firstly, to derive the reliable prototype for class k, we present an adaptive rebiased prototype, which simultaneously explore the global information and local information encoded in D_l. Specifically, the global information is embedded in matrix z_k where the embeddings of all samples in the training dataset belonging to class k are concatenated, whereas local information is collected in set x_k which consists of the embeddings of samples in current training batch belonging to category k. Based on z_k and x_k, we formulate the adaptive rebiased prototype for class k:

$$q_k = \frac{1}{|x_k|} \sum_{\mathbf{x}_{k,j} \in x_k} \frac{z_{k,j} z_k^\top}{\|z_{k,j}\| \, \|z_k\|} \tag{1}$$

where $|x_k|$ is number of samples in category k, $z_{k,j} = \Phi_{pro}(\mathcal{F}(\mathrm{x}_{k,j}))$ and $||\cdot||$ denotes l_2 norm.

Once we obtain the prototype, the distance between the representation of unlabeled sample i, z_i^u, and the prototype corresponding to class k, q_k, can be measured, which can be formulated as $d_{i,k} = Dist(z_i^u, q_k)$. There are various choices for the distance function $Dist(\cdot)$, and Euclidean distance [39] is selected here.

Afterwards, taking into account that known classes of unlabeled data are the same as the categories of labeled data, an unlabeled sample x_i^u belonging to k-th class shows an exponential decay with its distance from the prototype q_k, that is, $\mathbb{P}(x_i^u \mid q_k) \propto e^{-\mathrm{Dist}(q_k, x_i^u)}$ [40], which exhibits sharp distributional characteristics due to its fast decaying properties. Gaussian Mixture Models (GMM) [41] offer flexibility in adjusting distribution sharpness. Thus, we can suppose that the aforementioned $d_{i,k}$ subject to a GMM with the number of unknown classes $|C^n|$.

$$p(d_{i,k}) = \sum_{k=1}^{|C^n|} \phi_k \mathcal{N}\left(d_{i,k} \mid \mu_k, \Sigma_k\right) \tag{2}$$

where ϕ_k denotes the k-th weight component. $\mathcal{N}\left(d_{i,k} \mid \mu_k, \Sigma_k\right)$ is the Gaussian distribution with mean μ_k and covariance Σ_k. In this way, GMM can assign a label for each sample by calculating the probability that a sample belongs to each Gaussian distribution, and then we can estimate the number of class k on unlabeled samples.

As the number of known/unknown classes on unlabeled data is obtained, we assign different thresholds for known and unknown classes to reduce the bias towards the known classes, facilitating the learning of the unknown classes. First, we compute the maximum classification confidence $\widehat{p}_i = \max(p(\mathcal{A}_s(x_i^u)))$ and pseudo-label $\widehat{\mathbf{y}}_i = \arg\max p(\mathcal{A}_w(x_i^u))$. Then, we introduce a metric U to quantify the difference in learning between known and unknown classes.

$$U = \frac{\frac{1}{N_{known}} \sum_{x_i \in \mathcal{X}_{known}} \widehat{p}_i}{\log(\frac{1}{N_{unknown}} \sum_{x_j \in \mathcal{X}_{unknown}} \widehat{p}_j) + \nu} \tag{3}$$

where N_{known} and $N_{unknown}$ represent the number of samples in known classes and unknown classes. $\mathcal{X}_{known}$ denotes examples whose pseudo-labels correspond to known classes and $\mathcal{X}_{unknown}$ denotes examples with pseudo-labels belong to unknown classes. ν is a small positive value, preventing the denominator from vanishing. Log transformation smoothes extreme confidence in unknown classes and enhances U stability.

Futhermore, to better learn unknown classes, we employ U to adaptively adjust the confidence threshold for unknown classes. Specifically, we use the confidence threshold τ for known classes [25], and then heuristically set the threshold $\tau - U$ for unknown classes. The probability distribution $\mathbb{P}(X_{filter})$ of the filtered set of samples can be expressed as follows:

$$\begin{aligned} \mathbb{P}(X_{filter}) = & \sum_{\mathbf{x}_i^u \in \mathcal{X}_{known}} \mathbb{I}(\widehat{p}_i \geq \tau) p(\mathcal{A}_s(\mathbf{x}_i^u)) \\ & + \sum_{\mathbf{x}_j^u \in \mathcal{X}_{unknown}} \mathbb{I}(\widehat{p}_j \geq \tau - U) p(\mathcal{A}_s(\mathbf{x}_j^u)) \end{aligned} \tag{4}$$

To encourage the model to produce a more balanced probability distribution of predictions, we incorporate an adaptive factor, denoted $w = log(\mathbb{P}(X_{filter}))$, to compute the cross-entropy between the predictions of the strong augmented examples and the corresponding pseudo-labels as an unsupervised loss:

$$\mathcal{L}_{un} = \sum_{\mathbf{x}_i^u \in \mathcal{X}_{known} \cup \mathcal{X}_{unknown}} \mathbb{I}(\widehat{p}_i \geq \tau_i) H(\widehat{\mathbf{y}}_i, p(\mathcal{A}_s(x_i^u)) + w) \tag{5}$$

where $p(\mathcal{A}_s(x_i^u))$ denotes prediction for sample i. $H(\cdot, \cdot)$ is the cross-entropy function. τ_i is τ for x_i belongs to $\mathcal{X}_{known}$, $\mathbb{I}(\cdot)$ is the indicator function.

To further increase the confidence of pseudo-labels for minority classes, we propose a post-processing scheme to fine-tune the unsupervised loss. Observations indicate that general models often misclassify minority class samples, justifying efforts to focus model training on these less frequent samples. However, without any intervention, given a sample belonging to a minor category, models always tend to categorize it to a major class, which is incorrect. Therefore, we utilize the posterior probability to design a coefficient r_i, enforcing models to pay more attention to samples of minor classes:

$$r_i = \frac{1}{\nu + T(1 - p(l|\mathcal{A}_s(x_i^u)) + R)} \tag{6}$$

where T and R are hyper-parameters. $p(l|\mathcal{A}_s(x_i^u))$ denotes the posterior probability of the sample i belonging to l-th class. We will experimentally show that these parameters are insensible to various datasets. Next, we can use the coefficient r_i to pay more attention to these samples from minor classes. Finally, we integrate it into the loss function to enhance the quality of pseudo-labels:

$$\mathcal{L}_{un} = \sum_{\mathbf{x}_i^u \in \mathcal{X}_{known} \cup \mathcal{X}_{unknown}} r_i \cdot \mathbb{I}(\widehat{p}_i \geq \tau_i) H(\widehat{\mathbf{y}}_i, p(\mathcal{A}_s(x_i^u)) + w) \tag{7}$$

In this manner, our SRE jointly optimizes the estimation of known or unknown classes and adaptive post-processing unsupervised loss to mitigate the bias of unknown classes towards majority known classes.

D. Calibrate Representation with Synergistic Refinement Scheme

To calibrate misconnections among fine-grained classes in graphs constructed from DKIE data representations, caused by structured annotations, we propose a **S**ynergistic **R**efinement (SR) mechanism. This mechanism is designed to accurately identify the boundaries and internal structure of entities in DKIE. To be more specific, a **D**ynamic **T**hreshold **H**ash **R**ectification (DTHR) mechanism is proposed to remove erroneous connections between samples of fine-grained classes. Besides, a **S**tructural **C**alibration (SC) mechanism is presented to identify the connected components with the same class label by largest connectivity component (LCC) on the graph. To further distinguish fine-grained class samples, we introduce multi-granularity graph contrastive learning for the graph.

1) Construct Consensus Graph: To explore the relationship of fine-grained classes on the unlabeled set, we employ $k-$NN [42] to construct the undirected graph, denoted as $\mathrm{G} = \langle \mathrm{V}, \mathrm{E} \rangle$. This correlation graph is constructed on their representation similarities. Here, V represents the set of graph vertices, and E denotes the edges connecting these vertices. Each edge $\mathbf{g}_a(i,j)$ of the correlation graph G_a is obtained by:

$$\mathbf{g}_a(i,j) = \begin{cases} (z_i^u)^\top \cdot \boldsymbol{z}_j^u, & \text{if } i \neq j \wedge \boldsymbol{z}_i^u \in \mathbf{NN}_k(z_j^u) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Here, z_i^u is the latent representation of unlabeled x_i^u, and NN_k is the k-neighborhood, the selection of k detailed in Table XIII.

2) Affinity Updating with Dynamic Threshold Hash: Locality-Sensitive Hashing (LSH) [43] widely used for similarity search in structured documents, following the principle that similar data points in original space maintain their similarity in binary hash space [44]. However, these methods with fixed thresholds fail to adjust to the specific characteristics of DKIE data: low thresholds might mislabel similar items, and high ones could group dissimilar items incorrectly.

To address this issue, we introduce a dynamic threshold hash that disconnects false connections between fine-grained classes by adaptively removing low confidence samples. Specifically, given the representation z^u of unlabeled samples, we define the hash matrix H as the result of the matrix multiplication of z_i^u and z_j^u. Each vector $h_i \in H$ is assigned a value based on a simple rule: If the dot product of representations z_i^u and z_j^u is positive, h_i is set to 1. Otherwise, it is set to 0. In this way, unlabeled representations are mapped into a hash matrix H, facilitating efficient similarity searches. Then, we set the edge of G_a to 1 if the Hamming distance between two hashes is below dynamic threshold, the generated binarized affinity graph G_b of G_a, formulated as:

$$\mathbf{G}_b(i,j) = \begin{cases} 1, & \text{if } d\left(h_i, h_j\right) \leq \xi \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where the hash values h_i and h_j comes from H, d denotes

Algorithm 1 The overall algorithm of our method.

Input: $\mathcal{D}_l = \left\{\left(x_i^l, y_i^l\right)\right\}_{i=1}^n$, unlabeled subset $\mathcal{D}_u = \{(x_i^u)\}_{i=1}^m$. $\mathcal{A}_w(\cdot)$ and $\mathcal{A}_s(\cdot)$: weak and strong augmentation functions. $\mathcal{F}(\cdot)$: Feature encoder. $\Phi_{cls}(\cdot)$: classifier. $\Phi_{pro}(\cdot)$: projection head C^u: set of all classes, T_{proto} : temperature factor
while not reach the maximum iteration **do**
  for $c = 1$ to step **do**
    $z^u = \Phi_{pro}(\mathcal{F}(\mathcal{A}_s(x^u)))$, $p_{cls} = \Phi_{cls}(z^u)$
    Compute class prototype q_k on labeled data by Eq. (1), A set of prototype $Q = \{q_k\}_{k=1}^{|C^u|}$
    Compute the distance between prototype q_k and each of $x_i^u \in \mathcal{D}_u$ by $d_{i,k} = Dist(z_i^u, q_k)$
    $\mu_k, \sigma_k^2 \leftarrow \arg\max \mathcal{N}\left(d_{i,k} \mid \mu_k, \sigma_k^2\right)$ // estimate classes on unlabeled data via GMM
    Compute the adaptive factor w using the metric U in Eq. (3) and Eq. (4) to compute the adaptive factor w.
    Compute the post-processing coefficient r_i by Eq. (6)
    Compute the unsupervised loss $\mathcal{L}_{un}$ by Eq. (7) with r_i and w
    Construct the k-NN graphs for feature embeddings z^s and z^w by Eq. (8)
    Construct the k-NN graphs for class embedddings p_s and p_{cls} by Eq. (8)
    Remove sample x_i and its adjacent edges from K-NN graph by Eq. (9), resulting in binary graph
    Remove small connected components of the binary graph, that is, only the largest connected components is retained, then update the binary graph by Eq. (12) and generate binarized affinity graph by Eq. (13), G^{ws}, G^{sw}, G^{ws}, G_{cls} and G_{cls}^q
    Compute the representation-level graph contrastive loss based on G^{ws} and G^{sw} by Eq. (14)
    Compute the class-level graph contrastive loss based on G_{cls} and G_{cls}^q by Eq. (15)
    The overall multi-granularity graph contrastive loss can be formulated by Eq. (16)
    $\mathcal{F}, \Phi_{cls}, \Phi_{pro} \leftarrow$ Adam with $\mathcal{L}_{\text{total}}$ formulated by $Eq.(16)$
  end for
end while
Return: The well trained $\mathcal{F}$, Φ_{cls}, Φ_{pro}.

TABLE I
THE DIVISION OF KNOWN AND UNKNOWN CLASSES ON FUNSD AND CORD.

Dataset	Known Classes	Unknown Classes
FUNSD	["menu.nm", "menu.price", "menu.cnt", "total.total_price", "sub_total.subtotal_price", "total.cashprice", "total.changeprice", "sub_total.tax_price", "menu.unitprice", "menu.sub_nm", "total.menuqty_cnt", "menu.discountprice", "total.creditcardprice", "sub_total.service_price", "sub_total.etc", "sub_total.discount_price", "menu.sub_cnt", "menu.sub_price", "total.emoneyprice", "total.menutype_cnt", "menu.num", "total.total_etc", "menu.etc"]	["menu.etc", "menu.sub_unitprice", "menu.sub_etc", "menu.vatyn", "menu.itemsubtotal", "sub_total.othersvc_price", "void_menu.nm", "void_menu.price"]
CORD	["I-ANSWER", "B-ANSWER", "I-QUESTION", "B-QUESTION", "Other"]	["I-HEADER", "B-HEADER"]

the Hamming distance. A smaller Hamming distance implies greater similarity. To better simulate the threshold, we convert similarity to an angle value and scale by a constant. The threshold ξ is adaptive, which could be computed by:

$$\xi = \hat{K} \cdot \frac{\arccos(\bar{s})}{\pi} \tag{10}$$

where $\hat{K}$ denotes the length of the hash value. $\bar{s}$ denotes average cosine similarity between normalized hash values in each batch.

$$\bar{s} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1, j \neq i}^{B} \frac{h_i \cdot h_j}{\|h_i\| \, \|h_j\|} \tag{11}$$

In this manner, we can dynamically update the nearest neighbor candidate points on the graph. Specifically, nodes with matching hash values will form edges, while unmatched samples and adjacent edges will be discarded.

3) Construct Sub-graph with Structural Calibration: For the l-th class unlabeled samples, smaller connected components often include samples from different classes on the graph G_b of fine-grained classes. To address this issue, we introduce a structure refinement mechanism, where the Largest Connected Component (LCC) can improve the quality of class-specific subgraph by calibrating the error connections among fine-grained classes. We select LCC sub-graphs for each class to refine structures.

$$G_c = \bigcup_{l=1}^{|\mathcal{C}^u|} G_b(l)_{LCC}. \tag{12}$$

Here, $G_b(l)_{lcc}$ is class-specific subgraph for l-class. G_c represents the set of LCCs of all categories.

Then, we build an affinity graph on corresponding correlation graph G_a, utilizing the set of Largest Connected Components (LCCs) G_c along with pseudo-labels. The generated binarized affinity graph G_d is formulated as:

$$\mathbf{G}_d(i,j) = \begin{cases} 1, & (y_i = y_j) \wedge (I_{LCC}(i,j)) \\ 0, & (y_i \neq y_j) \end{cases} \tag{13}$$

where y_i and y_j are pseudo-labels. $I_{LCC}(i,j)$ is an indicator function, that evaluates to 1 if the indices i and j correspond to nodes that are in the same G_c. In binarized affinity graph G_d, positive/negative pairs are treated as credible pseudo-positives/negatives for enhancing graph contrastive learning at multi-granularity.

4) Multi-granularity Graph Contrastive Learning: To further refine the connectivity on both known and unknown classes, we design multi-granularity contrast learning on representation-level and class-level subgraphs, both of which are performed simultaneously. First, for feature embedding, inspired by [45], we employ two augmented representations $z^w = \Phi_{pro}(\mathcal{F}(\mathcal{A}_w(x^u)))$ and $z^s = \Phi_{pro}(\mathcal{F}(\mathcal{A}_s(x^u)))$ of unlabeled data with a randomized way to obtain a blended representation z^{ws} as query, denoted as $z^{ws} = \alpha z^s + (1-\alpha) z^w$, hyperparameter α is discussed in Section IV-C. To maintain the shift-invariance, we define the corresponding key z^{sw} of z^{ws}defined as $z^{sw} = \alpha z^w + (1-\alpha) z^s$. Second, for class embedding, we integrate semantic similarity [46] p_s into the predicted probabilities p_{cls} to calibrate the bias in the predicted probabilities, which can be denoted as $\hat{p}_{cls} = p_{cls} \cdot p_s$. Meanwhile, we utilize cross-multiplication to compute the query of p^q_{cls} to improve the consistencies, formulated as $\hat{p}^q_{cls} = p_s \cdot p_{cls}$. Then, we use k-NN to construct correlation graphs for the class and feature embeddings, then generate binary affinity graphs through the SR mechanism (detailed in Section D(1)-(3)). Finally, we introduce graph contrastive learning for class-level and representation-level.

$$\mathcal{L}_p^{con} = \frac{1}{m} \sum_{i=1}^{m} \log \frac{\exp\left(\mathrm{sim}\left(U_i^{ws}, U_i^{sw}\right)\right)}{\sum_{i'=1, i' \neq i}^{m} \exp\left(\mathrm{sim}\left(U_i^{ws}, U_{i'}^{sw}\right)\right)} \tag{14}$$

$$\mathcal{L}_{cls}^{con} = \frac{1}{m} \sum_{i=1}^{m} \log \frac{\exp\left(\mathrm{sim}\left(\hat{V}_i, \hat{V}_i^q\right)\right)}{\sum_{i'=1, i' \neq i}^{m} \exp\left(\mathrm{sim}\left(\hat{V}_i, \hat{V}_{i'}^q\right)\right)} \tag{15}$$

where $U_i^{ws} = MLP\left(G_i^{ws}\right)$, $U_i^{sw} = MLP\left(G_i^{sw}\right)$, $\hat{V}_i = MLP\left(G_{cls,i}\right)$ and $\hat{V}_i^q = MLP\left(G_{cls,i}^q\right)$, MLP denotes another projection head [47], $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. G^{ws} and G^{sw} are the affinity graphs corresponding to z^{ws} and z^{sw}, respectively. G_{cls} and G^q_{cls} are the affinity graphs corresponding to $\hat{p}_{cls}$ and $\hat{p}^q_{cls}$, respectively. The overall multi-granularity graph contrastive loss is formulated as:

$$\mathcal{L}_{mul} = \gamma \mathcal{L}_{cls}^{con} + (1-\gamma) \mathcal{L}_p^{con} \tag{16}$$

where γ=0.5 serves as the trade-off parameter, detailed in Section IV-C.

E. Overall Optimization Objective

In this study, SRE jointly optimizes three losses: (1) supervised classification loss L_{sup} on the labeled sample, (2)

TABLE II
EVALUATION RESULTS OF COMPARED METHODS ON KNOWN, UNKNOWN AND ALL CLASSES ON TWO DKIE BENCHMARKS. BOLD MEANS A STATE-OF-THE-ART SSL FRAMEWORK. IN OUR EXPERIMENTS, WE REPORT THE MEAN F1-SCORE AND STANDARD DEVIATION FROM FIVE RUNS TO REDUCE RANDOMNESS EFFECTS. "FULLY-SUPERVISED" REFERS TO TRAINING WITH ALL TRAINING EXAMPLES. DAGGER(†) DENOTES THE ORIGINAL METHOD CANNOT DETECT/RECOGNIZE UNKNOWN CLASSES (AND WE HAD TO EXTEND IT).

Method	CORD			FUNSD		
	Known	Unknown	All	Known	Unknown	All
Fully-supervised	94.38±1.33	63.57±1.02	94.12±1.21	93.03±0.79	63.88±0.53	90.71±0.45
RankStats [31]	77.54±0.63	31.08±0.59	73.87±0.41	42.43±0.67	21.07±0.35	41.08±0.46
UNO [32]	80.68±0.68	33.76±0.45	77.16±0.48	43.46±0.14	23.16±0.36	42.21±0.27
CKD [33]	81.74±0.37	35.37±0.46	78.37±0.48	43.76±0.24	26.14±0.43	43.56±0.34
†T2T [35]	82.59±0.37	36.78±0.21	79.38±0.71	42.75±0.34	23.46±0.56	41.67±0.54
†OpenMatch [13]	82.63±0.37	39.72±0.46	80.13±0.65	43.56±0.32	25.14±0.76	42.43±0.45
†SAFESTUDENT [37]	83.88±0.62	40.74±0.53	81.25±0.72	44.39±0.58	28.81±0.44	43.45±0.37
†IOMatch [38]	84.38±0.43	41.09±0.45	82.01±0.72	45.37±0.63	29.07±0.52	44.41±0.43
†FixMatch [25]	79.77±0.83	32.76±1.03	78.85±0.65	41.24±0.42	20.18±0.37	39.78±0.44
†SoftMatch [48]	83.99±0.52	41.76±0.45	81.66±0.57	46.12±0.48	26.20±0.31	44.87±0.66
†SimMatch [45]	82.56±0.55	42.33±0.43	82.18±0.37	43.78±0.32	24.32±0.42	42.63±0.37
†FreeMatch [7]	84.16±0.87	42.30±0.32	82.07±0.29	46.25±0.27	27.97±0.14	44.67±0.16
†DARP [9]	77.89±0.42	33.12±0.53	74.43±0.55	42.89±0.35	22.33±0.40	41.36±0.46
†CReST [10]	81.43±0.32	34.53±0.45	79.07±0.47	44.23±0.34	24.22±0.51	42.87±0.42
†ACR [30]	82.44±0.42	35.47±0.37	78.56±0.52	45.06±0.24	25.07±0.43	43.58±0.48
†Adsh [4]	82.29±0.37	35.30±0.38	78.66±0.35	45.42±0.34	25.31±0.63	43.82±0.49
ORCA [13]	84.12±0.63	43.56±0.52	81.78±0.43	45.66±0.57	30.27±0.45	44.39±0.48
GCD [14]	83.79±0.63	45.03±0.43	81.71±0.35	46.38±0.57	32.13±0.45	45.52±0.48
MetaGCD [11]	84.32±0.41	44.78±1.56	82.14±0.65	47.12±0.36	33.43±0.32	46.24±0.52
SRE (Ours)	**86.17±0.31**	**48.15±0.32**	**84.35±0.26**	**62.14±0.32**	**40.34±0.16**	**60.82±0.22**

unsupervised classification loss L_{un} on the unlabeled sample, and (3) multi-granularity graph contrastive Learning L_{mul} on unlabeled data.

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \mathcal{L}_{un} + \beta\mathcal{L}_{mul} \tag{17}$$

where β is the balance factor, detailed in Section IV-C. Following SimMatch [45], L_{sup} denotes cross-entropy loss.

$$\mathcal{L}_{sup} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{H}\left(y_b^l, \bar{p}\right) \tag{18}$$

where $\bar{p}$ is the prediction and y_b^l is label on the labeled set.

IV. EXPERIMENTS AND RESULTS

A. Experiment Details

Datasets: In our experiments, CORD and FUNSD are employed to evaluate the performance of the proposed SRE in DKIE. CORD [49] is released with 30 labels, which includes 800/100/100 receipts for training/validation/testing. FUNSD [1] comprises 199 receipts, labeled with 7 categories for 9,707 entities, split into 149 training and 50 test samples. For the known classes of each dataset, we randomly sample 5% of each class as labeled data and the rest as unlabeled data. Regarding the unknown classes, we select all data as unlabeled data since the small amount of data in these classes. We select the most minority classes as unknown classes and the rest as known classes. On FUNSD, "I-HEADER" and "B-HEADER" are unknown classes, the remaining 5 classes are known classes. On CORD, "menu.etc", "menu.sub_unitprice", "menu.sub_etc", "menu.vatyn", "menu.itemsubtotal", "sub_total.othersvc_price", "void_menu.nm", "void_menu.price" are unknown classes, the remaining 22 classes are known (detailed in Table I).

Implementations: Our experiments were carried out with NVIDIA Tesla V100 GPU. To demonstrate the effectiveness of our proposed SRE framework, all of the baselines use LayoutLMv3 as the base model to extract the features of inputs. We apply the Adam optimizer with a learning rate of 2e-5 and a weight decay of 1e-2. We adopt a batch size of 8 and train the networks for 80 epochs, with a linear learning rate warmup for the first 15 epochs. After the warmup, cosine scheduler is utilized. In our work, we use SimMatch [45] as baseline. F1-score as evaluation metric for FUNSD and CORD. In the augmentation strategy, we employ the Random Insertion [50] strategy as a strong augmentation $\mathcal{A}_s$ and Synonym Replacement [50] as weak augmentation $\mathcal{A}_w$. In this study, we set β = 0.2 (Section IV-C) and ν = 1e-5.
Baselines We compare SRE with SSL, long-tailed SSL, open-set SSL, NCD. Traditional SSL, long-tailed SSL and open-set SSL methods cannot discover unknown classes. To extend them to CMSS DKIE scenarios, we utilize SSL/open-set SSL to classify known classes and detect out-of-distribution (OOD) samples, Then, we use K-means to cluster OOD samples into unknown classes, allowing evaluation of their performance on unknown classes. We focus on the latest SOTA, including FixMatch [25], SoftMatch [48], SimMatch [45], and FreeMatch [7]. For long-tailed SSL, we report recent works such as DARP [9], CReST [10], Adsh [12] and ACR [30]. For open-set SSL, we consider the rencent works, including T2T [35], OpenMatch [13], SAFESTUDENT [37] and IOMatch [38]. For NCD, we select several latest SOTA methods: UNO [32], RankStats [31] and CKD [33]. Additionally, we introduce

TABLE III
ABLATION STUDY OF SRE ON FUNSD AND CORD. "PPE", "DTHR", "SC" AND "ML" MEAN LABELED-UNLABELED GRAPH, DYNAMIC THRESHOLD HASH RECTIFICATION, STRUCTURAL CALIBRATION MECHANISM AND MULTI-GRANULARITY GRAPH CONTRASTIVE LEARNING, RESPECTIVELY. "KNN" DENOTES k-NN GRAPH.

Index	Component					FUNSD			CORD		
	PPE	KNN	DTHR	SC	ML	Known	Unknown	All	Known	Unknown	All
1)						82.56	42.33	80.18	43.78	24.43	42.63
2)	✓					83.58	43.88	81.35	48.56	31.43	47.54
3)		✓			✓	82.79	43.75	80.63	50.23	33.27	49.23
4)		✓	✓		✓	83.16	45.57	81.26	52.37	34.61	51.32
5)		✓		✓	✓	82.75	44.85	81.02	54.56	36.26	53.47
6)		✓	✓	✓	✓	84.58	47.09	82.81	57.84	38.02	56.67
7)	✓	✓	✓	✓	✓	**86.17**	**48.15**	**84.35**	**62.14**	**40.34**	**60.82**

TABLE IV
ABLATION STUDIES OF SRE ON FUNSD AND CORD. DIFFERENT CLUSTERING ALGORITHMS, DIFFERENT PROTOTYPE ESTIMATION METHODS, AND DIFFERENT SSL METHODS ON OUR METHOD. ARP DENOTES ADAPTIVE REBIASED PROTOTYPE.

Data	Clustering Algorithm			Prototype		SSL methods		
	K-means	DBCAN	Ours (GMM)	Fixed Prototype	Ours (ARP)	FixMatch	SoftMatch	Ours (SimMatch)
CORD	78.77	80.12	81.35	80.56	81.35	80.03	80.88	81.35
FUNSD	42.43	43.06	47.54	43.36	47.54	43.44	43.96	47.54

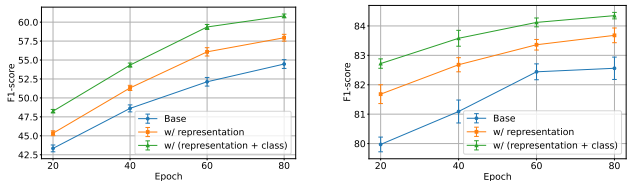
ORCA [13], GCD [14] and MataGCD [11] for comparison.

B. Main Results

Table II shows comparative results on the FUNSD and CORD datasets, including the F1-score of known classes, unknown classes, and all classes. From Table II, it can be observed that our approach can manifest substantial advantages over other baselines for both known and unknown classes. In particular, SRE outperforms previous SOTA (MetaGCD) by 15.02% in FUNSD and 1.85% CORD in known classes, while also obtaining 6.91% and 3.37% in unknown classes. It also remarkably surpasses MetaGCD by 14.58% on FUNSD and 20.9% on CORD in all classes. Traditional SSL, open-set SSL, and NCD methods struggle with the CMSS DKIE setting, but our method successfully addresses this challenge, outperforming baseline approaches. Specifically, compared to traditional SSL methods, SRE improves the F1-score by 15.89%-21.04% and 2.28%-5.5% on FUNSD and CORD in all classes. Additionally, long-tailed SSL methods assume that labeled and unlabeled data are from seen classes, which prevents them from adequately addressing the issue of unknown classes belonging to minority classes. In addition, the simple aggregation of long-tailed SSL and label correction method cannot work well in CMSS DKIE (shown in Table XII), since they fail to learn effective decision boundaries to distinguish between fine-grained classes. Our SRE outperforms open-set methods, improving the F1-score by 7.06%-16.88% for unknown classes and 2.34%-19.15% over all classes on two datasets, highlighting the value of utilizing unlabeled examples with unknown classes. Compared with NCD methods, our SRE clearly demonstrates the importance of jointly learning unknown classes during the training phase.

C. Ablation Study

Effectiveness of the Progressive Prototype Estimation. From Table III, we can see that PPE module can improve performance. Specifically, compared to baseline, PPE improves F1-score by 4.91% and 1.17% on FUNSD and CORD



(a) (b)

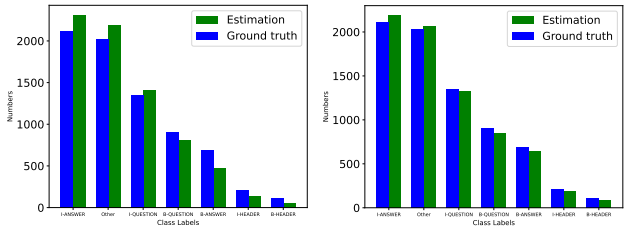
Fig. 3. The results of representation-level and class-level graph contrastive learning on the FUNSD (**a**) and CORD (**b**).

in all classes, respectively. This improvement highlights the ability of PPE module to mitigate the effects of the model's imbalance between majority and minority classes. In Fig. 4, the results show that the post-processing scheme contributes to obtaining an effective estimation of classes on unlabeled data. Additionally, we also compare GMM with other methods, such as DBSCAN [51] and K-means [52]. The results are reported in Table IV. From Table IV, we can see that GMM can obtain the best performance. Furthermore, we corroborate the efficacy of the adaptive rebiased prototype, as illustrated in Table IV. This indicates that PPE module can obtain better clustering results with reliable prototypes on labeled data. Moreover, from 6^{th} row of Table III, we can see that better prototypes can facilitate representation learning.

Effectiveness of the Synergistic Refinement Mechanism. We conduct ablation experiments on the SR mechanism in Table III. From 3^{rd} row, we can see that the performance of using KNN alone to perform graph contrastive learning shows a slight improvement over the baseline, attributed to insufficient understanding of fine-grained entity labeling. However, when we integrate the DTHR into the graph, there is a notable improvement in performance, with a 3.24%-10.18% increase in F1-score for unknown classes and a 1.08%-8.69% increase for all classes. This improvement proves that it is critical to reduce the harm caused by false negatives in contrastive loss by removing error connections for fine-grained classes. From 5^{th} row, the performance of the structural calibration mechanism can be further improved. This demonstrates the importance of counteracting the detrimental effects of false negatives in contrastive learning by retrieving more reliable positives. Meanwhile, from the 6^{th} row, we can see that dynamic threshold hash and structural calibration mechanism complement each other, revealing that this synergy facilitates refined connections among unknown classes and between unknown and known classes. To verify the impact of different components of multi-granularity graph contrastive loss, experiments with different level graph contrastive learning are conducted in Fig. 3. Representation-level has a greater impact than class-level, and this two-level contrastive loss interact with each other to jointly evolve.

TABLE V
THE TRAINING AND INFERENCE TIME ON CORD.

Methods	Training time (GPU Seconds)	Inference time (GPU Seconds)
FixMatch	358.88	0.448
SoftMatch	381.94	0.476
FreeMatch	327.55	0.410
SRE	335.31	0.419



(a) w/o PPS (b) w/ PPS

Fig. 4. The influence of post-processing scheme (PPS) on the FUNSD. denotes post-processing.

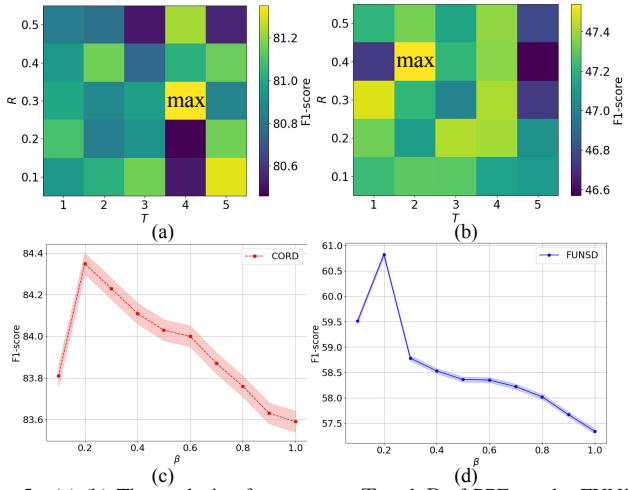


Fig. 5. (a)-(b) The analysis of parameters T and R of PPE on the FUNSD and CORD. (c)-(d) The analysis of factor β on the FUNSD and CORD.

Different Number of Labeled Data on SRE. We conduct the experiments using 10% and 20% labeled data. In our experiments, increasing the proportion of labeled data significantly enhance the model's performance. For instance, when the amount of labeled data was increased from 5% to 10%, we observed a 9.07% increase in F1-score for unknown classes and a 3.31% increase for all classes on the FUNSD. Further increasing the labeled data to 20% led to a 6.82% increase in F1-score for unknown classes and a 15.64% increase for all classes on the FUNSD. For CORD, the performance gains with increase in labeled data. These findings underscore the importance of labeled data in improving model generalization.

Influence of different numbers of unknown classes. We conducted ablation studies of the number of different unknown classes at CORD on SRE, as shown in Table X. From Table X, we can see that with the count of unknown classes rises from 8 to 15 (half of the total classes), we observe an increase in their performance. However, when this number reaches 22, performance begins to decline. This indicates that the model's effectiveness in handling unknown classes is influenced by their quantity, improving up to a point where they constitute less than half of the total classes.

Evaluation with the unknown number of unknown classes. Our SRE and other baselines assume that number of unknown classes is known. For case that the number of unknown classes is not known a priori, we can first use technique proposed in [53] to estimate the number of clusters. We then retest baselines using the estimated class numbers. Specifically, on CORD dataset with 30 classes, we first apply technique presented in [53] to estimate the number of classes to be 35, and then we can apply SRE by estimating the number of classes. The results in Table XI show that SRE outperforms baselines such as UNO, SAFESTUDENT, FreeMatch and MetaGCD, with a 7.68% improvement over MetaGCD in unknown class. Moreover, with the estimated number of classes, SRE shows a relatively large gap in results compared to the setting in which the number of classes is known a priori.

TABLE VI
EVALUATION RESULTS OF DIFFERENT NUMBERS OF LABELED DATA ON SRE.

Method	FUNSD			CORD		
	Known	Unknown	All	Known	Unknown	All
5% labeled data	62.14	40.34	60.82	86.17	48.15	84.35
10% labelled data	71.21	43.45	69.65	88.12	52.87	86.55
20% labeled data	78.30	47.16	76.46	90.47	56.74	89.42

TABLE VII
ANALYSIS ON F1-SCORE OF COMBINATION BETWEEN EXISTING LONG-TAILED METHODS [9], [10], [12] AND INCORRECT LABEL CONNECTIONS STRATEGIES (E.G., METAGCD [11]). NOTE THAT WE SELECT METAGCD SINCE IT OUTPERFORMS OTHER SIMILAR MODELS.

Methods	Known	Unknown	All
MetaGCD	84.32	44.78	82.14
MetaGCN + DARP	84.76	44.81	82.55
MetaGCN + CReST	84.58	44.83	82.39
MetaGCN + Adsh	84.78	44.85	82.57
SRE	86.17	48.15	84.74

Analysis of Training and Inference Time. Different semi-supervised learning methods may improve performance by introducing additional training strategies, loss functions, or using unlabeled data, and these modifications tend to affect training time and inference time. We report the training time of different methods on CORD dataset with 800 data in Tabel IV-C. The GPU seconds metric is calculated based on the V100 GPU. Despite achieving slightly worse results in training time and inference time compared to SoftMatch, our approach achieves a significant performance improvement and achieves SOTA. So, our approach achieves a trade-off between performance and time cost.

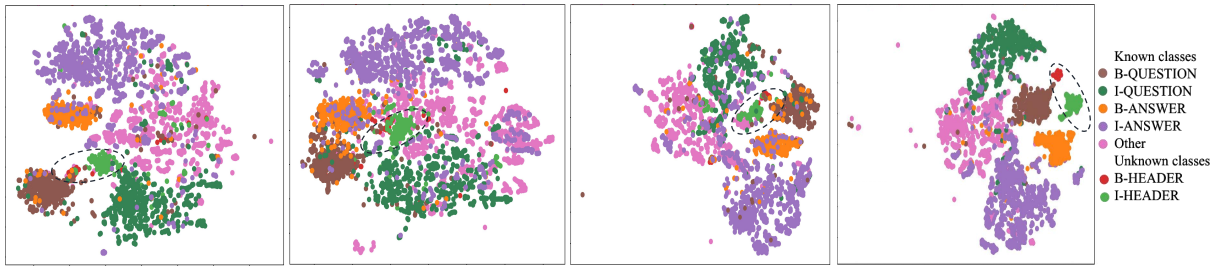
TABLE VIII
ABLATION STUDIES OF DIFFERENT α IN SRE ON THE FUNSD AND CORD.

Dataset	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
CORD	83.43	83.53	83.74	84.08	84.11	84.35	82.43	82.84
FUNSD	57.62	57.84	58.65	58.76	59.77	60.82	58.43	57.06

TABLE IX
ABLATION STUDIES OF DIFFERENT γ IN EQ. 16 ON THE FUNSD AND CORD.

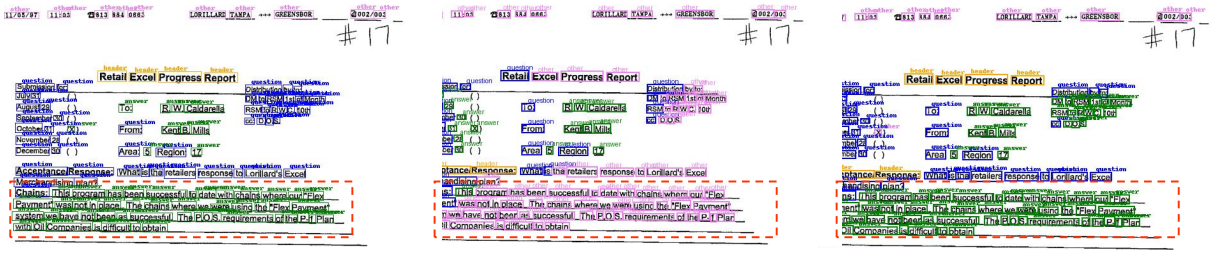
Dataset	1	2	3	4	5	6	7	8
CORD	83.43	83.53	83.74	84.08	84.35	83.21	81.33	81.33
FUNSD	57.53	57.99	58.76	58.55	60.82	59.42	58.64	57.21

Parameter Sensitivity Analysis. We evaluate the influence of β in Eq. (17) on the FUNSD and CORD datasets. Results (all classes) are presented in Fig. 5-(c-d). The results demonstrate that our SRE can achieve the best performance when $\beta = 0.2$. Additionally, we analysis the parameters T and R of PPE,



(a) FreeMatch (b) SAFESTUDENT (c) GCD (d) SRE

Fig. 6. T-SNE visualization on FUNSD.



(1) (2) (3)

Fig. 7. Examples visualization of SRE on the CORD datasets: (1) Ground truth, (2) prediction of W/O unlabeled examples that contain unknown classes, (3) result of W/ unlabeled examples that contain unknown classes. Entities denoted with blue, violet, green, **orange** for Known class, Known class, Known class and Unknown class, respectively.

depicted in Fig. 5. From Fig. 5-(a-b), PPE can achieve the best performance with $T = 2$ and $R = 0.4$ for FUNSD and when $T = 4$ and $R = 0.3$ for CORD. Table VIII shows that $\alpha = 0.6$ (in Section III-D4) is the best performance. We set k=5 for k-NN, as illustrated in Table XIII. Table IX shows the selection of γ in Eq. 16.

Visualization of the Representation Space. Fig. 6 shows latent representation space of different methods on the FUNSD. Fig. 6-(a)-(b) reveals mixed class representations in FreeMatch and SAFESTUDENT without clear clustering. Fig. 6-(c) shows that GCN obtain better separation of unknown classes, yet with irregular shapes. In contrast, Fig. 6-(d) shows the latent space of our SRE. Here, a greater number of samples are closely grouped around the pertinent prototypes, significantly enhancing class distinction and minimizing bias within our PPE. Fig. 7 shows the influence of unlabeled examples with unknown classes.

TABLE X
ABLATION STUDIES OF THE NUMBER OF DIFFERENT UNKNOWN CLASSES AT CORD ON SRE.

Numbers of Unknown classes	Known	Unknown	All
8	86.17	48.15	84.35
15	87.87	49.65	85.13
22	85.17	45.21	76.29

TABLE XI
ABLATION STUDIES OF WITH UNKNOWN NUMBER OF UNKNOWN CLASSES ON CORD.

Methods	Known	Unknown	All
UNO	52.34	26.32	50.74
SAFESTUDENT	56.47	27.49	54.61
FreeMatch	54.58	25.61	52.63
MetaGCD	57.25	29.48	55.63
SRE	70.27	37.16	68.43

TABLE XII
ANALYSIS ON F1-SCORE OF COMBINATION BETWEEN EXISTING LONG-TAILED METHODS [9], [10], [12] AND INCORRECT LABEL CONNECTIONS STRATEGIES (E.G., METAGCD [11]). NOTE THAT WE SELECT METAGCD SINCE IT OUTPERFORMS OTHER SIMILAR MODELS.

Methods	Known	Unknown	All
MetaGCD	84.32	44.78	82.14
MetaGCN + DARP	84.76	44.81	82.55
MetaGCN + CReST	84.58	44.83	82.39
MetaGCN + Adsh	84.78	44.85	82.57
SRE	86.17	48.15	84.74

TABLE XIII
ABLATION STUDIES OF DIFFERENT k IN k-NN ON THE FUNSD AND CORD.

Dataset	1	2	3	4	5	6	7	8
CORD	83.43	83.53	83.74	84.08	84.35	83.21	81.33	81.33
FUNSD	57.61	57.85	58.80	59.07	60.82	58.46	58.23	57.11

Additional Ablation Study From Table XII, we can see that simple aggregation strategies cannot work well in CMSS task. In contrast, SRE improves the F1-score of unknown classes while sustaining the F1-score of the known classes.

V. CONCLUSION

The paper studies the class-missing semi-supervised DKIE problem. To this end, we present a new approach termed SRE, which is designed to address a key challenge: misconnections between known and unknown classes. Specifically, we design a progressive prototype estimation method to alleviate minority unknown class bias towards the majority known class on unlabeled data. Besides, we introduce a synergistic refinement mechanism for graphs constructed from latent representations to calibrate the connections among fine-grained classes. Extensive experiments and ablation studies have demonstrated that the proposed SRE surpasses other existing state-of-the-art methods on two public DKIE benchmarks. In the future,

we will address this by performing risk analyses on unseen classes and expanding the proposed method to other tasks.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61973245 and the Key RD Plan of Shaanxi Province (Program No.2023-YBGY-029).

REFERENCES

[1] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2. IEEE, 2019, pp. 1–6.

[2] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," in *International Conference on Machine Learning*. PMLR, 2023, pp. 18 893–18 912.

[3] Y. Yu, Y. Li, C. Zhang, X. Zhang, Z. Guo, X. Qin, K. Yao, J. Han, E. Ding, and J. Wang, "Structextv2: Masked visual-textual prediction for document image pre-training," in *International Conference on Learning Representations (ICLR)*, 2023.

[4] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1192–1200.

[5] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4083–4091.

[6] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[7] Y. Wang, H. Chen, Q. Heng, W. Hou, M. Savvides, T. Shinozaki, B. Raj, Z. Wu, and J. Wang, "Freematch: Self-adaptive thresholding for semi-supervised learning," *arXiv preprint arXiv:2205.07246*, 2022.

[8] Y. Yang, D.-C. Zhan, Y.-F. Wu, Z.-B. Liu, H. Xiong, and Y. Jiang, "Semi-supervised multi-modal clustering and classification with incomplete modalities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 2, pp. 682–695, 2019.

[9] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 14 567–14 579, 2020.

[10] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 857–10 866.

[11] Y. Wu, Z. Chi, Y. Wang, and S. Feng, "Metagcd: Learning to continually learn in generalized category discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1655–1665.

[12] L.-Z. Guo and Y.-F. Li, "Class-imbalanced semi-supervised learning with adaptive thresholding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8082–8094.

[13] K. Cao, M. Brbic, and J. Leskovec, "Open-world semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2022.

[14] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7492–7501.

[15] T. I. Denk and C. Reisswig, "Bertgrid: Contextualized embedding for 2d document representation and understanding," *arXiv preprint arXiv:1909.04948*, 2019.

[16] M. Kerroumi, O. Sayem, and A. Shabou, "Visualwordgrid: Information extraction from scanned documents using a multimodal approach," in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 389–402.

[17] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, "Pick: processing key information extraction from documents using improved graph learning-convolutional networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4363–4370.

[18] Y. Qian, E. Santus, Z. Jin, J. Guo, and R. Barzilay, "Graphie: A graph-based framework for information extraction," *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)*, 2019.

[19] G. Tang, L. Xie, L. Jin, J. Wang, J. Chen, Z. Xu, Q. Wang, Y. Wu, and H. Li, "Matchvie: Exploiting match relevancy between entities for visual information extraction," pp. 1039–1045, 2021.

[20] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.

[21] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che *et al.*, "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding," pp. 2579–2591, 2021.

[22] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8934–8954, 2022.

[23] Y. Yang, Z.-Y. Fu, D.-C. Zhan, Z.-B. Liu, and Y. Jiang, "Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 2, pp. 696–709, 2019.

[24] Y. Yang, D.-W. Zhou, D.-C. Zhan, H. Xiong, Y. Jiang, and J. Yang, "Cost-effective incremental deep model: Matching model capacity with the least sampling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3575–3588, 2021.

[25] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[26] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[27] Z. Hu, Z. Yang, X. Hu, and R. Nevatia, "Simple: similar pseudo label exploitation for semi-supervised classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 099–15 108.

[28] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[29] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[30] T. Wei and K. Gan, "Towards realistic long-tailed semi-supervised learning: Consistency is all you need," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3469–3478.

[31] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "Automatically discovering and learning new visual categories with ranking statistics," *arXiv preprint arXiv:2002.05714*, 2020.

[32] E. Fini, E. Sangineto, S. Lathuiliere, Z. Zhong, M. Nabi, and E. Ricci, "A unified objective for novel class discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9284–9292.

[33] P. Gu, C. Zhang, R. Xu, and X. He, "Class-relation knowledge distillation for novel class discovery," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 16 428–16 437.

[34] L.-Z. Guo, Z.-Y. Zhang, Y. Jiang, Y.-F. Li, and Z.-H. Zhou, "Safe deep semi-supervised learning for unseen-class unlabeled data," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3897–3906.

[35] J. Huang, C. Fang, W. Chen, Z. Chai, X. Wei, P. Wei, L. Lin, and G. Li, "Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8310–8319.

[36] K. Saito, D. Kim, and K. Saenko, "Openmatch: Open-set semi-supervised learning with open-set consistency regularization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 956–25 967, 2021.

[37] R. He, Z. Han, X. Lu, and Y. Yin, "Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 585–14 594.

[38] Z. Li, L. Qi, Y. Shi, and Y. Gao, "Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization," in

Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15 870–15 879.

[39] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[40] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," *Advances in neural information processing systems*, vol. 17, 2004.

[41] H. Permuter, J. Francos, and I. Jermyn, "A study of gaussian mixture models of color and texture features for image classification and segmentation," *Pattern recognition*, vol. 39, no. 4, pp. 695–706, 2006.

[42] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer, 2003, pp. 986–996.

[43] A. Gionis, P. Indyk, R. Motwani *et al.*, "Similarity search in high dimensions via hashing," in *Vldb*, vol. 99, no. 6, 1999, pp. 518–529.

[44] B. Stein, "Principles of hash-based text retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 527–534.

[45] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu, "Simmatch: Semi-supervised learning with similarity matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 471–14 481.

[46] Y. Oh, D.-J. Kim, and I. S. Kweon, "Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9786–9796.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[48] H. Chen, R. Tao, Y. Fan, Y. Wang, M. Savvides, J. Wang, B. Raj, X. Xie, and B. Schiele, "Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning," in *Eleventh International Conference on Learning Representations*. OpenReview. net, 2023.

[49] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee, "Cord: a consolidated receipt dataset for post-ocr parsing," in *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[50] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6382–6388.

[51] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE, 2014, pp. 232–238.

[52] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[53] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8401–8409.

[54] P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, and H. Liu, "Selfdoc: Self-supervised document representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5652–5660.

[55] Y. Xu, L. Shang, J. Ye, Q. Qian, Y.-F. Li, B. Sun, H. Li, and R. Jin, "Dash: Semi-supervised learning with dynamic thresholding," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 525–11 536.

[56] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[57] Y. Chen, X. Zhu, W. Li, and S. Gong, "Semi-supervised learning under class distribution mismatch," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3569–3576.