Learning Geometry-Aware Representations for New Intent Discovery

Kai Tang[1], Junbo Zhao[1], Xiao Ding[2], Runze Wu[3],
Lei Feng[4], Gang Chen[1], Haobo Wang[1]*,
[1]Zhejiang University [2]Harbin Institute of Technology
[3]NetEase Fuxi AI Lab [4]Singapore University of Technology and Design
{tk0819,j.zhao,cg,wanghaobo}@zju.edu.cn,xding@ir.hit.edu.cn
wurunze1@corp.netease.com,lfengqaq@gmail.com

Abstract

New intent discovery (NID) is an important problem for deploying practical dialogue systems, which trains intent classifiers on a semi-supervised corpus where unlabeled user utterances contain both known and novel intents. Most existing NID algorithms place hope on the sample similarity to cluster unlabeled corpus to known or new samples. Lacking supervision on new intents, we experimentally find the intent classifier fails to fully distinguish new intents since they tend to assemble into intertwined centers. To address this problem, we propose a novel **GeoID** framework that learns geometry-aware representations to maximally separate all intents. Specifically, we are motivated by the recent findings on Neural Collapse (NC) in classification tasks to derive optimal intent center structure. Meanwhile, we devise a dual pseudo-labeling strategy based on optimal transport assignments and semi-supervised clustering, ensuring proper utterances-to-center arrangement. Extensive results show that our GeoID method establishes a new state-of-the-art performance, achieving a $+\mathbf{3.49}\%$ average accuracy improvement on three standardized benchmarking datasets. We also verify its usefulness in assisting large language models for improved in-context performance. The code is available at https://github.com/zjutangk/GeoID.

1 Introduction

Intent detection, which aims at recognizing the intention of the user query, is an important component in industrial-grade dialogue systems (Qin et al., 2021; Li et al., 2022). Though there has been significant progress in recent years (Perkins and Yang, 2019; Min et al., 2020; Vedula et al., 2020), the effectiveness of these models greatly depends on the predefined intent labels, which often fall short of fulfilling the application requirements in an open-world environment. To address this problem, new intent discovery (NID) has been proposed that learns from a semi-supervised corpus containing: (i)-a limited amount of utterance annotated with true intents; (ii)-a vast quantity of unlabeled data containing both known intents and unknown new intents. The goal of NID is to simultaneously classify known intents while discovering those unknown intent clusters.

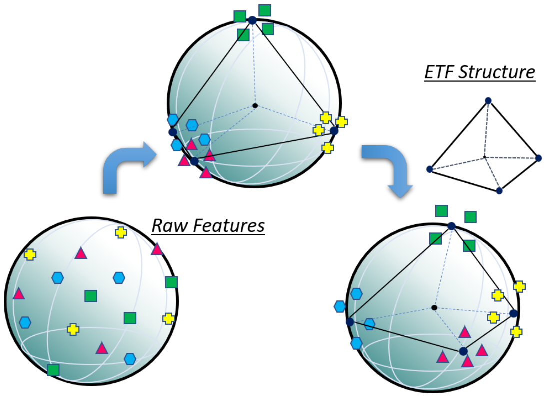


Figure 1: Illustration for the effect of our off-line ETF structure: Having gotten suboptimal features from the simple similarity-based method, we achieve an optimal feature distribution structure by bringing the samples closer to a fixed ETF structure.

Emerging attempts have been made to tackle the NID task (Zhou et al., 2023a; Mou et al., 2023; Wang et al., 2023). Amongst them, the most popular methods (Zhou et al., 2023b,c) rely on the similarity-based assumption, i.e., samples sharing the same intent typically located at a neighbor region. Then, they design specific objectives to spontaneously group different samples into separated clusters, e.g. pairwise constraints (Lin et al., 2020), contrastive learning (Zhang et al., 2022), and graph diffusion (Shi et al., 2023). However, without new intent supervision, the learning procedure can be dominated by the labeled samples, resulting in biased feature representations. Thus, the unknown

[*] Corresponding author.

samples are prone to collapse into intertwined centers, thus being inseparable, as empirically verified in Figure 5. This gives rise to a representation dilemma—*how to obtain the optimal hypersphere distribution for identifying new intents?*

To answer this question, we draw inspiration from a recent study on the neural collapse (NC) phenomenon (Papyan et al., 2020). It implies that, under ideal conditions, as a classifier undergoes training towards convergence, the final-layer characteristics of a specific class gradually converge towards their average value within that class. These averages then tend to position themselves on a hypersphere with the utmost angular distance, known as the simplex equiangular tight frame (simplex ETF), as referenced in (Martinez and Kak, 2001). The NC phenomenon provides an optimal frame of hypersphere centers that aligns with our expectations for the final intent centers to be maximally separated. Nevertheless, it is non-trivial to achieve NC for NID with largely missing supervision.

In this paper, we propose a novel **Geo**metry-aware representation learning framework for **NID** (dubbed **GeoID**), which encourages both known/unknown intents to be *equiangular separative* and *maximally discriminative*. Specifically, GeoID replaces the learnable weights of the intent classifier (implicitly the class centers) with fixed ETF weights. That is, it pulls the sample features to predefined intent centers instead of merely clustering by sample similarity themselves. Moreover, we develop a dual pseudo-labeling strategy to ensure proper sample allocation to the centers: (i)-optimal transport-based objective to refine pseudo-labels from classifier predictions, which is reliable with pattern fitting; (ii)-semi-supervised clustering-driven pseudo-labeling that is more robust in the early stages. Finally, We train the NID classifier to fit both pseudo-labels for improved robustness.

We comprehensively evaluate GeoID on three NID benchmark datasets. Specifically, our visualized analysis shows that GeoID does indeed learn highly distinguishable intent clusters (Figure 5(c)), and its mean intent centers almost achieve neural collapsed distribution (Figure 3(c)). Thanks to this excellent geometry property, our GeoID establishes new state-of-the-art NID performance, e.g., it improves the accuracy by $+\mathbf{2.90}\%$, $+\mathbf{3.88}\%$, and $+\mathbf{3.69}\%$ accuracy on the BANKING, StackOverflow, and CLINC datasets, respectively. We hope our work will inspire future NID studies to tackle this important problem of cluster separability.

2 Related Work

New Intent Discovery (NID). NID is a developing research area aiming to identify unlabeled utterances from known and new intents (Wu et al.; Liang and Liao, 2023; Kumar et al., 2022; An et al., 2023a). To address this issue, initial efforts were made from the perspective of constrained clustering. Hsu et al. (2018a,b) utilized pairwise similarity-constrained clustering and refinement module to discover new intents and Zhang et al. (2021) employed pretraining techniques and aligned clustering labels to learn clustering-friendly representations. Subsequent work aims to learn more discriminative feature representations from the perspective of contrastive learning. Wei et al. (2022) emphasized intra-class compactness by employing contrastive learning following (Kim et al., 2021; Yan et al., 2021; Giorgi et al., 2021) to bring samples with the same pseudo-label closer in feature space. An et al. (2023b) adopted a decoupled training approach to construct prototype learning, aiming to obtain more discriminative features. Zhang et al. (2022) harnessed neighbor contrastive learning and more challenging pre-training tasks to achieve outstanding performance. In addition, Zhang et al. (2023a) integrated techniques from previous works including aligned clustering and contrastive learning. These works primarily tackle the intra-class compactness issue while neglecting the influence of bias on inter-class separation.

Neural Collapse (NC). The NC phenomenon was initially discovered by Papyan et al. (2020). They observed that towards the end of training, a classification model would exhibit the collapse of last-layer features towards their respective within-class centers. These within-class centers, along with the classifiers, will converge to form a simplex equiangular tight frame (ETF). Subsequent studies have sought to provide theoretical insights into this elegant phenomenon. It has been proven that neural collapse corresponds to the global optimality of simplified models under various conditions, including regularization (Tirer and Bruna, 2022; Zhou et al., 2022; Zhu et al., 2021)and constraints (Fang et al., 2021; Graf et al., 2021). These findings hold for both cross-entropy (CE) (Ji et al., 2021; Graf et al., 2021) and mean squared error (MSE) loss functions (Han et al., 2021; Zhou et al., 2022). Phenomenon of neural collapse (NC) has also been explored in specific scenarios, such as imbalanced learning (Xie et al., 2023; Yang et al., 2022; Zhong et al., 2023), learning with noisy labels (Nguyen et al., 2022), and transfer learning (Galanti et al., 2021; Xiao et al., 2024), but they have never tapped in language domain. We are the first to investigate the NC phenomenon in the NID task.

3 Method

Problem Statement. Confronted with a new intent discovery task, we are usually given an intent dataset $\mathcal{D}_{\text{train}}$ encompassing two subsets: a labeled set $\mathcal{D}_l = \{(x_i, y_i) \mid y_i \in \mathcal{C}_{known}\}_{i=1}^{m}$, and an unlabelled set $\mathcal{D}_u = \{x_i \mid y_i \in \mathcal{C}_{known} \cup \mathcal{C}_{novel}\}_{i=1}^{n}$, where x_i refers to the input utterance,$\mathcal{C}_{known}$ is the set of known intent labels and $\mathcal{C}_{novel}$ is the set of new intent labels. We set $|\mathcal{C}_{known} \cup \mathcal{C}_{novel}|$ to L. NID can be viewed as a direct extension of general category discovery and the objective comprises two main goals: firstly, to identify new intents from unlabelled data, and secondly, to accurately classify inputs into their respective intents.

Intent Representation. Similar to the majority of mainstream tasks, we use BERT (Devlin et al., 2019) to extract intent representations. First, we input utterance x_i to BERT and get all its token embeddings $[\text{CLS}, T_1, \dots, T_M]$. Next, we utilize mean-pooling and dense layer to obtain intent feature representation $\boldsymbol{z}_i$. For contrastive learning, we utilize Random Token Replacement (RTR) following (Zhang et al., 2022) as data augmentation strategy to generate $\boldsymbol{z}_i{}'$.

Overview of Our Approach. Figure 2 illustrates the overall architecture of our method. We learn geometry-aware representations to maximally separate all intents by pushing utterances close to the corresponding vertex of the ETF structure inspired by the NC phenomenon. To achieve this, we devise pseudo-labeling a dual strategy based on optimal transport and clustering to assign samples to the correct centers.

3.1 Neural Collapse for Separation

In this section, we introduce the neural collapse (NC) phenomenon (Papyan et al., 2020) to demonstrate the characteristics of an ideal intent classifier for NID. In concrete, NC says the features learned from deep neural networks, in which last-layer features have the following appearances:(1) For each class, features collapse to the class mean. (2) The within-class means of all classes are located on a hypersphere and form a simplex equiangular tight frame (Simplex ETF).

Simplex ETF. *Suppose the vector space is d-dimensional. When $d \geq L - 1$, we can always derive a collection of L equal-length and maximally-equiangular d-dimensional embedding vectors $\boldsymbol{E} = [\boldsymbol{e}_1^*, \dots, \boldsymbol{e}_L^*] \in \mathbb{R}^{d \times L}$ to construct a simplex equiangular tight frame (ETF).*

$$\boldsymbol{E} = \sqrt{\frac{L}{L-1}}\mathrm{U}(\boldsymbol{I}_L - \frac{1}{L}\mathbf{1}_L\mathbf{1}_L^\top) \tag{1}$$

where $\boldsymbol{I}_L$ is the identity matrix, $\mathbf{1}_L$ is a vector of all ones and $\mathrm{U} \in \mathbb{R}^{d \times L}$ is a rotation matrix.

As mentioned above, an optimal ETF structure for the arrangements of features facilitates the minimization of within-class variance and the maximization of between-class variance for the features. This is very appealing for the NID task to distinguish all known/unknown intents.

In our GeoID framework, we calculate a set of pre-assigned centers $\boldsymbol{E} = [\boldsymbol{e}_1^*, \dots, \boldsymbol{e}_l^*] \in \mathbb{R}^{d \times L}$, each of which is a vertex of a random simple ETF structure according to Eq. (1). Then we optimize the following cross-entropy loss to pull these samples close to these centers:

$$\mathcal{L}_{etf}(\boldsymbol{x}_i, \hat{y}_i) = -\log \frac{\exp(\boldsymbol{z}_i^\top \cdot \boldsymbol{e}_{\hat{y}_i}^*)}{\sum_{l=1}^{L} \exp(\boldsymbol{z}_i^\top \cdot \boldsymbol{e}_l^*)} \tag{2}$$

where $\boldsymbol{z}_i$ is feature representation of x_i gotten by deep neural model. The key difference here from traditional CE loss is that we keep the classifier weights $\boldsymbol{E}$ fixed.

There remains one crucial problem — how to assign a sample $\boldsymbol{x}_i$ to its corresponding ETF vertex $\boldsymbol{e}_{\hat{y}_i}^*$. For labeled utterances (x_i, y_i), we can directly set ground-truth label y_i as $\hat{y}_i$. For the rest of unlabeled utterances, we resort to a pseudo-labeling algorithm to group them to proper centers, as described in the sequel.

3.2 Dual Pseudo-Labeling

To promote the formation of the novel class NC structure, we design two complementary pseudo-label generation strategies to guide the samples moving toward their corresponding ETF vertices.

Optimal Transport-based Strategy. Having gotten the softmax-based classifier prediction $\boldsymbol{P} = \{\boldsymbol{p}_i \mid \boldsymbol{p}_i = \text{softmax}(\boldsymbol{z}_i^\top \cdot \boldsymbol{E})\}$, one straightforward technique to obtain pseudo-labels is to adopt the maximum index of prediction (Sohn et al., 2020). However, due to the lack of supervision information, the classifier can be largely dominated by

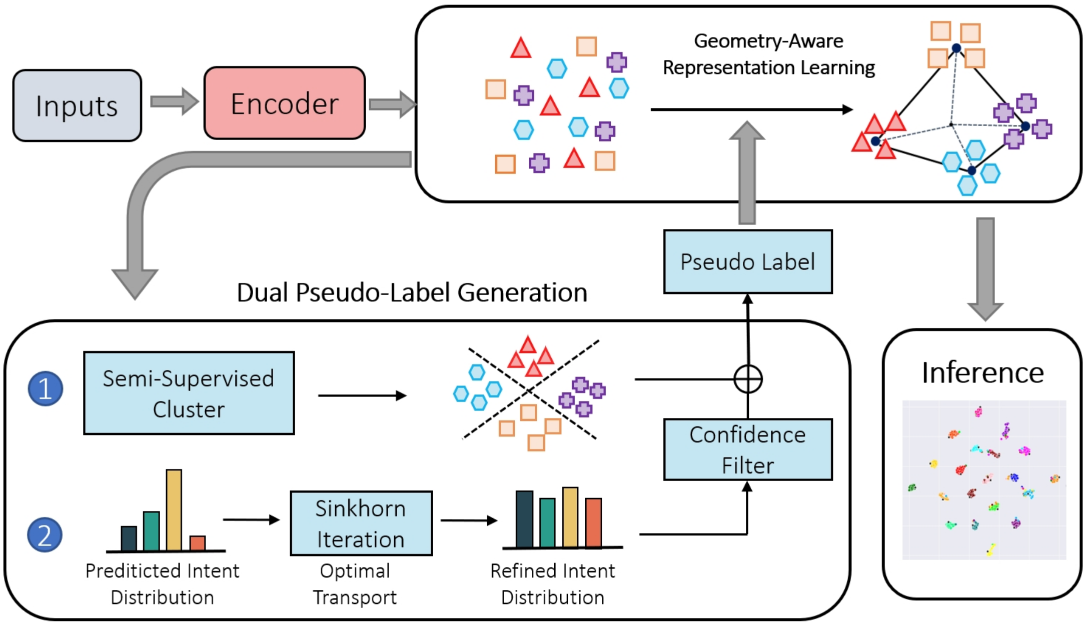


Figure 2: Overall framework of GeoID. We use an ETF structure to guide samples clustering into a pre-defined optimal distribution. To ensure the appropriate allocation of samples to the centers, we design a dual pseudo-labeling strategy based on optimal transport assignments and semi-supervised clustering. These components jointly promote learning geometry-aware representation.

known classes, making the representation of unknown classes collapse. To mitigate this problem, we introduce equipartition constraint into label assignment and transform pseudo-labeling into an optimal transport problem. Formally, given a batch of b unlabelled samples, we search for optimal label assignment $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_b]$ close to the current prediction $\boldsymbol{P}$ while subject to some constraints:

$$\boldsymbol{Q} = \max_{\boldsymbol{Q}\in\Delta} \mathrm{Tr}(\boldsymbol{Q}^\top \boldsymbol{P}) = \sum_{i=1}^{i=b}\sum_{j=1}^{j=L} q_{ij}^\top p_{ij} \qquad (3)$$

$$\text{s.t. } \Delta = \{[q_{ij}]_{L\times b} \mid \boldsymbol{Q}\mathbf{1}_b = \boldsymbol{r}, \boldsymbol{Q}^\top \mathbf{1}_L = \boldsymbol{c}\}$$

where $\mathrm{Tr}(\cdot)$ is the trace function. q_{ij} is the j-th element of pseudo-label assignment for the i-th sample. $\boldsymbol{r} = \frac{1}{L}\mathbf{1}_L$ is an L-dimensional uniform probability distribution indicating the prior class distribution. We set $\boldsymbol{c} = \frac{1}{b}\mathbf{1}_b$ indicating that b samples in a batch are sampled uniformly. The solution of Eq.(3) can be obtained by the Sinkhorn-Knopp algorithm (Cuturi, 2013); see Appendix A.3.

Such a strong assignment inevitably introduces a significant amount of noise. To alleviate the accumulation of label errors, we filter pseudo-labels by high-confidence selection. For each class $j \in [1, L]$, we select samples$\mathcal{D}_{sel}^j$:

$$\mathcal{D}_{sel}^j = \{(\boldsymbol{x}_i, y_i^{ot}) \mid y^{ot} = j, q_{ij} > \tau_j\} \qquad (4)$$

where $y_i^{ot} = \arg_j \max q_{ij}$ is the predicted category label of the ETF classifier on $\boldsymbol{x}_i$. τ_j is the confidence threshold that enables selecting the top $R\%$ ranked samples in the data subset whose prediction is class j. Lastly, we only calculate $\mathcal{L}_{ETF}$ on $\mathcal{D}_l$ and $\mathcal{D}_{sel} = \cup_{j=1}^{L} \mathcal{D}_{sel}^j$. In other words, we select the most reliable samples within each cluster and employ them as pivots to guide the whole cluster being moved toward the ETF centers.

Clustering-based Strategy. Even with optimal transport-based allocation, the classifier can be unreliable at the early stage of training. To this end, we supplement a data-driven strategy by using feature-based clustering to regularize the learning procedure. It should be noted such a data-driven strategy does not suffer from the bias caused by the absence of supervision and thus largely enhances the robustness of GeoID during the early stages.

Specifically, we harness the rationality of clustering algorithms to generate pseudo-labels once per epoch. Considering that we have labeled data available, we adopt semi-supervised k-means following (Vaze et al., 2022) as the clustering algorithm. The indices of clustering labels obtained by k-means are randomly permuted in each training epoch which makes it challenging for us to leverage the supervision information provided by the clustering results fully. To address this issue, we devise a clustering label alignment strategy that combines with simplex ETF structure. To be specific, we align the cluster centers of different epochs with the pre-assigned centers $\boldsymbol{E}$ provided by NC to ensure consistent cluster assignment in semi-supervised k-means employing Hungarian algorithm (Kuhn, 2010) and obtain the aligned cluster label y^{align} which can be found in Appendix A.4.

Finally, we allow the model to fit both pseudo-labels in a trade-off manner :

$$\mathcal{L}_{cls} = \alpha\mathcal{L}_{etf}(\boldsymbol{x}, \boldsymbol{y}^{ot}) + (1-\alpha)\mathcal{L}_{etf}(\boldsymbol{x}, \boldsymbol{y}^{align}) \tag{5}$$

where $\boldsymbol{y}^{ot}$ and $\boldsymbol{y}^{align}$ respectively correspond to the set of pseudo-labels generated by the two strategies. α is the trade-off hyperparameter ramped up from 0 to 1 during the training process.

In the beginning, our ETF classifier tends to trust the clustering pseudo-labels to avoid representation collapse. As the training proceeds, the filtered OT labels can gradually become more trustworthy, guiding the data clusters to move toward predefined ETF frames. In the meantime, the data-driven clustering pseudo-label still serves a proper regularization term that aligns the remaining data to its filtered neighbors, ensuring full data utilization. Finally, all data samples are tightly clustered to their corresponding centers, achieving the *optimal geometry distribution* we desire.

3.3 Contrastive Representation Enhancement

Finally, we also involve contrastive learning for improved representation. To explore semantically neighboring information and further enhance intra-class compactness and inter-class variance, we propose a contrastive loss following (Khosla et al., 2020) which brings similar samples close to improve clustering performance. Specifically, the positive sample set $\mathcal{P}_i$ for sample x_i from two aspects: (1) samples sharing the same ground-truth labels or pseudo labels and (2) k-nearest neighboring samples using dot product as a distance metric.

$$\mathcal{P}_i = \{j \mid z_j \in k\text{-NN}(\boldsymbol{z}_i) \vee y_j = y_i\} \tag{6}$$

where y_i is respectively grand-truth label for labeled sample and y^{ot} for unlabeled sample in $\mathcal{D}_{sel}$. Different from (Zhang et al., 2022) which conducts contrastive learning within a mini-batch, we mine neighbor samples from the entire dataset and utilize self-supervision signals in addition.

$$\mathcal{L}_{con} = -\frac{1}{|\mathcal{P}_i|}\log\frac{\sum_{j\in\mathcal{P}_i}\exp(\boldsymbol{z}_i^\top \boldsymbol{z}_j)}{\sum_{k=1}^{n}\exp(\boldsymbol{z}_i^\top \boldsymbol{z}_k)} \tag{7}$$

Dataset	$\mid \mathcal{C}_{known} \mid$	$\mid \mathcal{C}_{novel} \mid$	$\mid \mathcal{D}_{train} \mid$	$\mid \mathcal{D}_{test} \mid$
BANKING	19	58	9,003	3,080
StackOverflow	5	15	18,000	1,000
CLINC	37	113	18,000	2,250

Table 1: Statistics of our datasets.

We also involve a conventional consistency regularization term $\mathcal{L}_{reg}$ for improved semi-supervised training, please refer to Appendix A.2 for more details. The overall training loss is given by,

$$\mathcal{L}_{all} = \mathcal{L}_{cls} + \mathcal{L}_{con} + \mathcal{L}_{reg} \tag{8}$$

4 Experiment

4.1 Datasets

We first evaluate our method on four benchmark intent datasets. **Banking** (Casanueva et al., 2020) is a fine-grained imbalanced intent classification dataset collected from banking dialogues. **StackOverflow** (Xu et al., 2015) is a large-scale question classification dataset collected from online queries. **CLINC** (Larson et al., 2019) is an intent classification dataset from multiple domains. For each dataset, we randomly select 25% of the classes as $\mathcal{C}_{known}$, while the remaining classes are treated as $\mathcal{C}_{novel}$. Detailed statistics of datasets are summarized in Table 1.

4.2 Baselines

We compare our method with various new intent discovery baselines, including CDAC+ (Lin et al., 2020), DeepAligned (Zhang et al., 2021), MTP-CLNN (Zhang et al., 2022), and UNISD (Zhang et al., 2023a). We also include comparisons with constrained clustering and novel class discovery methods: KCL (Hsu et al., 2018a), MCL (Hsu et al., 2018b), GCD (Vaze et al., 2022), and DTC (Han et al., 2019). For clustering and novel class discovery methods used for computer vision tasks, we adapt them for our task by leveraging the BERT backbone. For performances of baselines, we directly adopt reported results from receptive papers or UNISD. For a fair comparison, we remove the use of the extern dataset in MTP-CLNN.

4.3 Evaluation Metrics

Following (Zhang et al., 2021, 2023a), We employ three commonly used metrics, namely Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Accuracy (ACC), to evaluate the quality of the clustering results. To calculate accuracy, we

Method	BANKING			StackOverflow			CLINC		
	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC
KCL	53.85	20.07	28.79	35.47	16.80	32.88	67.98	24.30	29.40
MCL	49.46	15.51	24.53	29.44	14.99	31.50	62.79	18.21	28.52
DTC	56.05	20.19	32.91	33.38	16.45	30.32	79.35	41.92	56.90
GCD	59.74	26.04	38.50	35.77	20.12	36.74	83.70	52.23	64.82
CDAC+	67.65	34.88	48.79	74.33	39.44	74.30	84.68	50.02	66.24
DeepAligned	69.85	37.16	49.67	53.97	36.46	53.96	88.97	64.63	74.07
MTP-CLNN	80.04	52.91	65.06	76.85	57.62	77.54	93.17	76.20	83.26
DWGF	79.04	51.17	61.40	73.68	56.59	75.00	93.65	78.38	84.44
UNISD	81.94	56.53	65.85	75.87	65.93	77.92	94.17	77.95	83.12
GeoID (ours)	**82.54**	**57.92**	**68.75**	**77.62**	**66.35**	**81.80**	**94.37**	**81.24**	**88.13**

Table 2: Performance on testing sets of different benchmarks. The labled sample ratio is set to 10%. Average results over 3 runs are reported. For each dataset, the best results are marked in bold.

utilize the Hungarian algorithm (Kuhn, 2010) to derive the optimal assignments between clustering classes and ground-truth classes.

4.4 Implementation Details

We use the pre-trained *bert-base-uncased* BERT model from (Wolf et al., 2019) as our backbone. In the pre-training phase, we employ the same settings following (Zhang et al., 2022). Regarding model optimization, We employ AdamW optimizer (Wolf et al., 2019) with a warm-up schedule and 0.01 weight decay. The learning rate is set to $1e^{-4}$ for all benchmark datasets. For Sinkhorn-Knopp, we set hyperparameters following (Fini et al., 2021) that $\epsilon = 0.05$ and run for 3 iterations. For mining k-nearest neighbors, we set $k = 64$ for BANKING and CLINC, $k = 256$ for StackOverflow. For pseudo-labeling, we set R of high-confidence selection as 30%, 10% in begin respectively for $\mathcal{C}_{known}$, $\mathcal{C}_{novel}$ and linearly increases to 80% in 70 epochs. Additionally, to enhance training efficiency with the BERT backbone, we freeze all transformer layer parameters except the last layer following (Lin et al., 2020). Following most of the baselines, we run cluster-based evaluation when testing. Details of consistency regularization can be found in Appendix A.2. Although we involved several hyperparameters, most of them followed the default settings, which had minimal impact on the experiments.

Method	NMI	ARI	ACC
GeoID	**82.54**	**57.92**	**68.75**
w/o Optimal Transport	79.48	53.42	65.21
w/o Clustering	81.34	55.92	66.10
w/o ETF	78.85	51.24	63.59

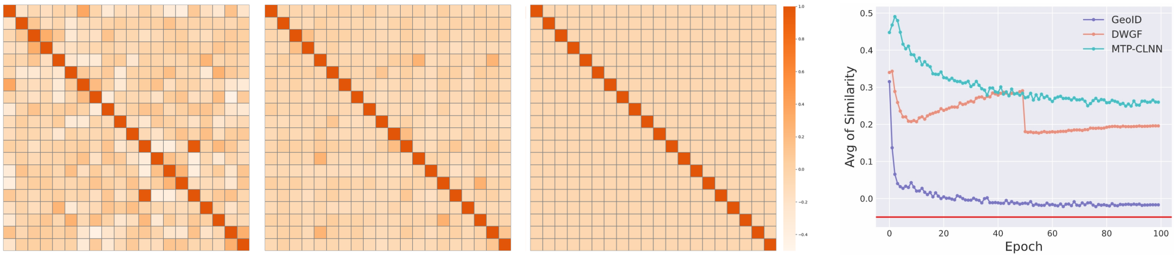
Table 3: Ablation study on BANKING dataset. The known class ratio and labeled ratio are respectively set to 0.25 and 0.1.

4.5 Main Results

Table 2 shows the main experimental results on the BANKING, StackOverflow, and CLINC datasets. Our method consistently outperforms all competing methods by a considerable margin on all metrics across various datasets. Specifically, without an extern dataset in pre-training, our method achieves improvements on three benchmark datasets with $+\mathbf{2.90}\%$, $+\mathbf{3.88}\%$, and $+\mathbf{3.69}\%$ compared with best baselines. Moreover, in the scenario where the CLINC dataset is utilized as an external dataset, our model can maintain advantages on the Banking and StackOverflow respectively and details of this experiment are shown in Appendix B.1.

4.6 Discussion

Ablation Study. To further analyze the contributions of different components in our method, we conduct three ablation studies here. First, to analyze the effect of our dual pseudo-labeling strategy, we separately removed the optimal transport and clustering components during the pseudo-labeling process. Then we exclude the use of the ETF structure by replacing the ETF structure with a linear classification head while keeping the form of cross-entropy loss unchanged. These experiments are all conducted on the BANKING dataset to evaluate their effectiveness. As shown in Table 3, removing



(a)

(b)

(c)

(d)

Figure 3: Analysis on Neural Collapse phenomenon. Figure 3(a)-3(c) is the visualization of pair-wise cosine similarity. We calculate the cosine similarity on class means (a) replacing the ETF structure with a linear classification head and (b) using offline simplex ETF structure. Figure 3(c) is the cosine similarity on vertices of an ETF structure, representing the ideal scenario corresponding to the optimal distribution of features. The darker the color, the higher the cosine similarity between the corresponding two categories. Figure 3(d) is the quantitative result about the maximal separation character of NC and the red horizontal line corresponds to the optimal case $-\frac{1}{L-1}$.

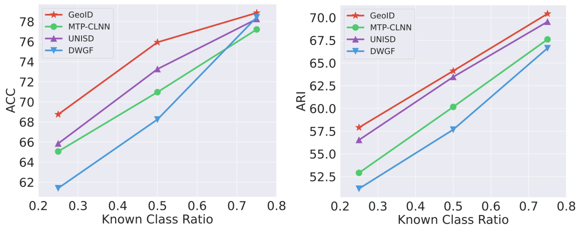


Figure 4: Comparisons on BANKING with varying known class ratios ranging from 0.25 to 0.75.

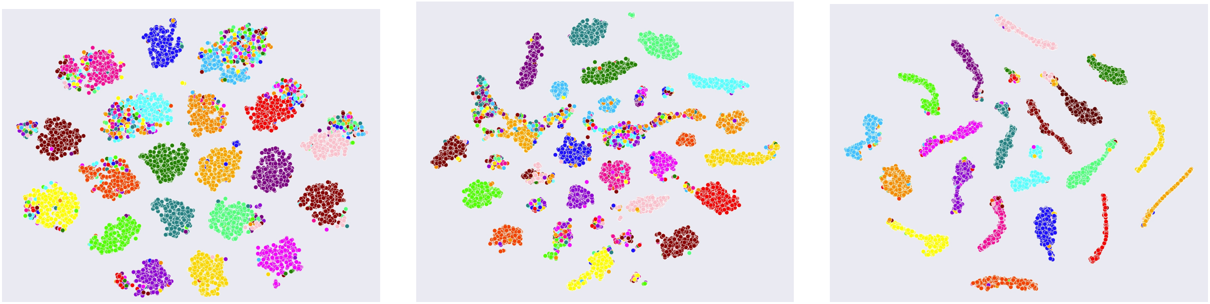
these components respectively brings negative effects of $-\mathbf{3.54}\%$, $-\mathbf{2.65}\%$, $-\mathbf{5.16}\%$ on accuracy, indicating that all these modules are beneficial to NID task. More details of the ablation study can be found in Appendix B.2.

Analysis on Neural Collapse. Considering the features of simplex ETF struct in Eq.(1), we can summarize the characteristics of the NC phenomenon from two perspectives: (1) uniform distribution on hypersphere (equiangularity) and (2) cluster centers are maximally separated. To further investigate the impact of NC on feature distribution, we visualize the cosine similarity between cluster centers that intuitively demonstrate the equiangularity of NC, and the results are shown in Figure 3(a)-3(c). To quantitatively analyze the second character of NC phenomenon, we calculate the average value $\text{Avg}_{l\neq l'}(\cos(\hat{c}^l, \hat{c}^{l'}))$, where $\hat{c}^l$ is the means of features from class $1 \leq l \leq L$. For the maximal separation character of NC, the optimal pair-wise angle $-\frac{1}{L-1}$ is derived from Eq.(1). The trends of this metric on the BANKING dataset with different methods are shown in Figure 3(d). According to the visualization and quantitative results shown in Figure 3, we have demonstrated that the distribution of features in GeoID can closely approximates the ideal scenario of the neural collapse phenomenon (NC).

Analysis on Representation Learning. To investigate the effectiveness of our method in geometry-aware representation learning, we visualize the representation learned by GeoID and two strong baselines on StackOverflow using t-SNE (Van der Maaten and Hinton, 2008) in Figure 5. It can be shown that MTP-CLNN and DWGF produce intertwined cluster centers. In contrast, our method is demonstrated to effectively avoid representation collapse and learn clearly separable clusters. These results demonstrate that GeoID is indeed able to learn optimal geometry distributions for NID.

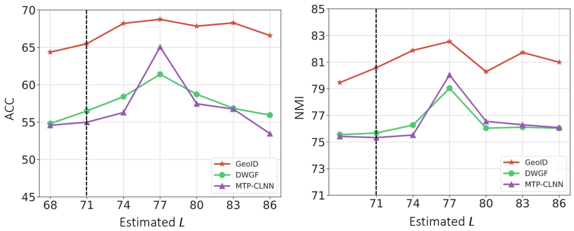
Influence of Known Class Ratio. Here we further evaluate the performance of our method on the BANKING dataset with different known class ratios from 0.25 to 0.75. As shown in Figure 4, our method consistently outperforms other baselines under varying known class ratios.

Unknown Number of Novel Classes. The number of all intents is an important hyperparameter in the NID task. The previous experiments were conducted based on the assumption that we already knew the number of novel classes which is often impractical in real-world applications. Now some methods to estimate class number have been proposed, such as (Han et al., 2019; Zhang et al., 2021). These methods can reliably maintain the error ratio below 10%, for example, methods in (Zhang et al., 2021) can estimate L in BANKING to be 71. So we conducted simulations on the BANKING dataset



(a) MTP-CLNN (b) DWGF (c) GeoID

Figure 5: T-SNE visualization of the representation distribution on StackOverflow. Different colors indicate the corresponding category of intents.



(a) ACC (b) NMI

Figure 6: Comparison result of ACC and NMI under different estimated number of labels. The black vertical line denotes the result under DeepAligned's estimation.

Method	**BANKING**		
	Known class	**Novel class**	**All Class**
ChatGPT-ZSD	52.72	51.24	51.99
ChatGPT-FSD	65.81	50.03	57.92
ChatGPT-GeoID	**67.29**	**67.55**	**67.42**

Table 4: Performance comparison of LLMs in different application forms. About ChatGPT-ZSD, we directly provide unlabeled samples to LLMs to cluster. In ChatGPT-FSD, we provide LLMs with labeled samples as prior knowledge.

using different prediction results of L to observe the impact of class number on model performance. In Figure 6, we show the performance variation of GeoID across different total numbers of predicted intent categories L. It can be observed that GeoID exhibits overall robustness across different total numbers of predicted intent categories.

Discussion of LLM. Recently, there have been explorations (Zhang et al., 2023b; Song et al., 2023) of applying large language models to the task of New Intent Discovery. In this section, we further investigate the feasibility of collaboration between LLMs and GeoID. Especially, we use GeoID to select corpora close to centers. This allows LLMs to perform in-context learning and subsequently perform direct clustering tasks. We design prompts following (Song et al., 2023) to implement two baselines ChatGPT-ZSD and ChatGPT-FSD for comparison. The results in Table 4 demonstrate that the supplement of novel class samples significantly improves the clustering performance of LLMs, particularly in balancing the performance gap between known classes and novel classes. Additionally, due to the limitations of LLMs in unsupervised clustering performance on corpora, their performance still falls short of fine-tuning SLMs. One may design other LLMs-based techniques specifically for clustering tasks for improved performance on NID. But, it is beyond the scope of our paper. More details and discussion can be found in appendix A.5.

5 Conclusion

In this paper, we propose a new method GeoID which revisits new intent discovery from the perspective of geometric-aware representation learning. Inspired by the neural collapse phenomenon, we use a fixed ETF structure to facilitate the learning of a feature distribution that aligns with the predefined optimal structure. Additionally, we design a dual pseudo-labeling strategy using optimal transport theory and semi-supervised clustering to generate pseudo-labels respectively. This strategy enhances the quality of pseudo-labels while maintaining sample utilization. Experiments show GeoID's significant improvements over baselines. We hope our work can inspire future research to further improve the NID problem from the perspective of geometry-aware representations.

Limitations

To inspire future research, we summarize the limitations of our method. Firstly, we do not design a more accurate method for category estimation, leaving ample scope for further enhancement. Secondly, we did not delve deeply into the collaboration between large and small models on NID tasks. Thirdly, we have conducted preliminary research on basic intent classification tasks within open-world scenarios, but there exist numerous other intricate data formats (like continual learning) that merit further exploration.

Acknowledgements

This work is supported by the Pioneer R&D Program of Zhejiang (No. 2024C01035), CCF-NetEase ThunderFire Innovation Research Funding (NO. 202305) and the Fundamental Research Funds for the Central Universities (No. 226-2024-00049).

References

Wenbin An, Feng Tian, Ping Chen, Qinghua Zheng, and Wei Ding. 2023a. New user intent discovery with robust pseudo label training and source domain joint-training. *IEEE Intelligent Systems*.

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. 2023b. Generalized category discovery with decoupled prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12527–12535.

Inigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. *ACL 2020*, page 38.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118.

Enrico Fini, Enver Sangineto, Stéphane Lathuiliere, Zhun Zhong, Moin Nabi, and Elisa Ricci. 2021. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292.

Tomer Galanti, András György, and Marcus Hutter. 2021. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. 2021. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409.

XY Han, Vardan Papyan, and David L Donoho. 2021. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2018a. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2018b. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. 2021. An unconstrained layer-peeled perspective on neural collapse. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540.

Harold W. Kuhn. 2010. The hungarian method for the assignment problem. In Michael Jünger, Thomas M. Liebling, Denis Naddef, George L. Nemhauser, William R. Pulleyblank, Gerhard Reinelt, Giovanni Rinaldi, and Laurence A. Wolsey, editors, *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*, pages 29–47. Springer.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1836–1853.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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)*, pages 1311–1316.

Yinfeng Li, Chen Gao, Xiaoyi Du, Huazhou Wei, Hengliang Luo, Depeng Jin, and Yong Li. 2022. Automatically discovering user consumption intents in meituan. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 3259–3269. ACM.

Jinggui Liang and Lizi Liao. 2023. Clusterprompt: Cluster semantic enhanced prompt learning for new intent discovery. Association for Computational Linguistics.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8360–8367.

Aleix M Martinez and Avinash C Kak. 2001. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233.

Qingkai Min, Libo Qin, Zhiyang Teng, Xiao Liu, and Yue Zhang. 2020. Dialogue state induction using neural latent variable models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3845–3852. ijcai.org.

Yutao Mou, Xiaoshuai Song, Keqing He, Chen Zeng, Pei Wang, Jingang Wang, Yunsen Xian, and Weiran Xu. 2023. Decoupling pseudo label disambiguation and representation learning for generalized intent discovery. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9661–9675. Association for Computational Linguistics.

Duc Anh Nguyen, Ron Levie, Julian Lienen, Eyke Hüllermeier, and Gitta Kutyniok. 2022. Memorization-dilation: Modeling neural collapse under noise. In *The Eleventh International Conference on Learning Representations*.

Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Hugh Perkins and Yi Yang. 2019. Dialog intent induction with deep multi-view clustering. 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, Hong Kong, China, November 3-7, 2019*, pages 4014–4023. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4577–4584. ijcai.org.

Wenkai Shi, Wenbin An, Feng Tian, Qinghua Zheng, Qianying Wang, and Ping Chen. 2023. A diffusion weighted graph framework for new intent discovery. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8033–8042.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10291–10304.

Tom Tirer and Joan Bruna. 2022. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, pages 21478–21505. PMLR.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2022. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2009–2020. ACM / IW3C2.

Pei Wang, Keqing He, Yutao Mou, Xiaoshuai Song, Yanan Wu, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. APP: adaptive prototypical pseudo-labeling for few-shot OOD detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3926–3939. Association for Computational Linguistics.

Feng Wei, Zhenbo Chen, Zhenghong Hao, Fengxin Yang, Hua Wei, Bing Han, and Sheng Guo. 2022. Semi-supervised clustering with contrastive learning for discovering new intents. *CoRR*, abs/2201.07604.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhenhe Wu, Xiaoguang Yu, Meng Chen, Liangqing Wu, Jiahao Ji, and Zhoujun Li. Enhancing new intent discovery via robust neighbor-based contrastive learning.

Ruixuan Xiao, Lei Feng, Kai Tang, Junbo Zhao, Yixuan Li, Gang Chen, and Haobo Wang. 2024. Targeted representation alignment for open-world semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang Xie, Yibo Yang, Deng Cai, and Xiaofei He. 2023. Neural collapse inspired attraction–repulsion-balanced loss for imbalanced learning. *Neurocomputing*, 527:60–70.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. *NAACL HLT 2015*, pages 62–69.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. 2022. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in Neural Information Processing Systems*, 35:37991–38002.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14365–14373.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023a. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023b. Clusterllm: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13903–13920. Association for Computational Linguistics.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2022. New intent discovery with pre-training and contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 256–269. Association for Computational Linguistics.

Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. 2023. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19550–19560.

Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. 2022. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pages 27179–27202. PMLR.

Yunhua Zhou, Jiawei Hong, and Xipeng Qiu. 2023a. Towards open environment intent prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2226–2240.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023b. A probabilistic framework for discovering new intents. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3771–3784.

Yunhua Zhou, Jianqiang Yang, Pengyu Wang, and Xipeng Qiu. 2023c. Two birds one stone: Dynamic ensemble for ood intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10659–10673.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. 2021. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834.

A More Details of Implement

A.1 Details of ETF Structure Initialization

We employ random initialization, a widely adopted and effective methodology in recent neural collapse research. As described in Eq.(1), we guarantee the randomness of the initialization by using a random rotation matrix $\mathbf{U}$. The procedure involves only two hyperparameters: intent number L and feature dimension d. So there are no extra hyperparameters in this process.

A.2 Details of Consistency Regularization

In practice, we use Random Token Replacement (RTR) following (Zhang et al., 2022) as our data augmentation strategy. We denote the augmented view of x as x'. Then we employ the r-drop method to obtain two representations, z_1' and z_2' for x' through the backbone network. Finally, we calculate the Kullback-Leibler divergence between z_1' and z_2' as $\mathcal{L}_{reg}$.

A.3 Sinkhorn-Knopp Algorithm

Formally, we define a cost matrix $\boldsymbol{M} = \exp(\frac{\boldsymbol{P}}{\varepsilon})$. The Sinkhorn-Knopp algorithm is conducted by:

$$\boldsymbol{Q} = \boldsymbol{M} \odot (\boldsymbol{\mu} \cdot \boldsymbol{v}^T) \tag{9}$$

where $\boldsymbol{u} \in \mathbb{R}^L$ and $\boldsymbol{v} \in \mathbb{R}^b$ are scaling coefficients vectors and are updated iteratively by,

$$\boldsymbol{\mu} \leftarrow \boldsymbol{c}./(\boldsymbol{M}\boldsymbol{v}), \boldsymbol{v} \leftarrow \boldsymbol{r}./(\boldsymbol{M}^\top \boldsymbol{\mu}) \tag{10}$$

where ./ denotes element-wise division. In practice, we utilize a small iteration number of 3.

A.4 Details of Alignment in Clustering

Having gotten pseudo-label assignment $y^{cluster}$ and class centers $\boldsymbol{C} = \boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_L$ from semi-supervised k-means. We obtain the optimal projection matrix $\boldsymbol{M}$ with Hungarian algorithm:

$$\boldsymbol{C} = \boldsymbol{M} \cdot \boldsymbol{E} \tag{11}$$

Then we get aligned pseudo labels:

$$\boldsymbol{y}^{align} = \boldsymbol{M}^{-1} \cdot \boldsymbol{y}^{cluster} \tag{12}$$

A.5 Details of LLMs Experiment

The main challenge of LLMs in NID task is the absence of novel classes prior which seriously affects the incontext learning of large models. To solve this problem, we use GeoID to select corpora close to the center to provide novel classes prior.

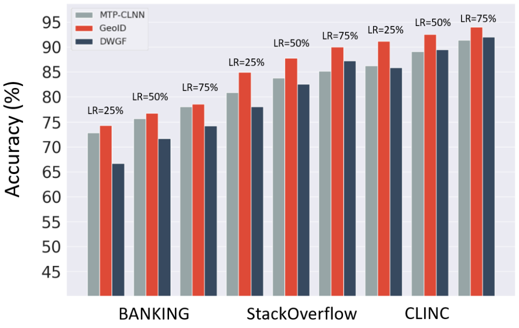


Figure 7: Accuracy comparison with different labeled ratios on different dataset.

In practice, for each novel class, we select 10 % samples nearest to the class center. We design the prompt in the form of <Prior: labeled samples and selected samples><Cluster Instruction><Response Format><D_{test}>.

B Additional Experimental Results

B.1 Pretraining with External Dataset

In the experimental section of the main text, to ensure a fair comparison, we excluded the external dataset used in the pretraining process of the MTP-CLNN (Zhang et al., 2022). In this section, we investigate the model's performance, which incorporates CLINC as the external dataset during the pretraining phase. The results shown in table 7 demonstrate that our method is equally capable of benefiting from the external dataset and maintaining a performance advantage on the benchmark.

B.2 More Ablation Study

In this section, we conducted additional ablation experiments regarding sample selection and contrastive learning. Regarding sample selection, we design two different schemes: without filtering and fixing the selection ratio as 50%. We also remove $\mathcal{L}_{con}$ during the training process to investigate the effect of contrastive learning. As shown in Table 6, removing them consistently impairs the model's performance, indicating that these designs indeed contribute to new intent discovery. We also add the ablation experiments on regularization.

B.3 Influence of Labeled Ratio

In addition to the known class ratio, we also analyze the influence of label ratio in model training. In the experiment, we vary the labeled ratio in the range of 25%, 50%, 75%. As shown in Table 7,

Method	GeoID	MTP-CLNN	DWGF	UNISD
Epoch Time	64.67s	68.62s	42.97s	68.92s

Table 5: Comparison of epoch time.

Method	NMI	ARI	ACC
GeoID	82.54	57.92	68.75
w/o filtering	81.02	54.31	65.74
fixed selection ratio	79.93	55.92	66.48
w/o $\mathcal{L}_{con}$	76.53	53.42	64.72
w/o $\mathcal{L}_{reg}$	81.29	56.14	68.02

Table 6: Additional ablation study on BANKING dataset. The known class ratio and labeled ratio are respectively set to 0.25 and 0.1.

our method achieves the best results with different labeled sample ratios.

B.4 More Fine-grained Metrics

In this section, we further analyze both the known class accuracy and unknown class accuracy to validate the effectiveness of the model on novel intents. As shown in Table 8, tt can be seen that the performance improvement of our algorithm mainly comes from the improved accuracy of novel intent samples. This is precisely realized based on the optimal distribution of features we obtained through the ETF structure.

B.5 Training Complexity

We compare the time required for the model to complete one training epoch with other works. As demonstrated in the table 5, GeoID is as fast as current state-of-the-art NID algorithms.

C Overall Algorithm

We summarize the pseudo-code of our proposed GeoID in Algorithm 1.

Algorithm 1: Pseudo-code of GeoID.

```
Input: Training dataset $\mathcal{D}$, model $f$, filter ratio $R$,
  num of intents $L$;
Initialize ETF strcuture $\boldsymbol{E}$
for $epoch = 1, 2, \ldots,$ do
    for $\boldsymbol{x}_i \in \mathcal{D}$ do
        Induce representation $\boldsymbol{z}_i = f(\boldsymbol{x}_i)$
        // optimal transport
        $\boldsymbol{p}_i = \boldsymbol{z}_i \cdot E$
        $\boldsymbol{q}_i =$ Sinkhorn($\boldsymbol{p}_i$) as Eq.(3)
    end
    // high-confidence selection
    $\mathcal{D}_{sel} =$ filter($\boldsymbol{q}$) as Eq.(4)
    $y^{ot} =$ argmax($q_{ij}$) for $x_i \in \mathcal{D}_{sel}$
    // cluster for pseudo-labeling
    obtain $y^{cluster}$ with $k$-means clustering($\boldsymbol{z}$)
    $y^{align} =$ Hungarian($y^{cluster}, \boldsymbol{E}$)
    calculate $\mathcal{L}_{cls}$ with ETF structure as Eq.(5)
    // overall training objectives
    minimize loss $\mathcal{L}_{all} = \mathcal{L}_{cls} + \mathcal{L}_{con} + \mathcal{L}_{reg}$
end
```

Method	**BANKING**			**StackOverflow**		
	NMI	**ARI**	**ACC**	**NMI**	**ARI**	**ACC**
DeepAligned	69.85	37.16	49.67	53.97	36.46	53.96
UNISD	81.94	56.53	65.85	75.87	65.93	77.92
MTP-CLNN (extern)	84.11	61.29	71.43	79.68	70.17	83.77
GeoID (extern)	**84.24**	**62.71**	**74.38**	**80.69**	**71.04**	**84.90**

Table 7: Performance on different benchmarks. The labeled sample ratio is set to 10% and the known class ratio is set to 25%. Average results over 3 runs are reported. For each dataset, the best results are marked in bold.

Method	**BANKING**		**StackOverflow**		**CLINC**	
	Known	**Novel**	**Known**	**Novel**	**Known**	**Novel**
DeepAligned	69.60	45.44	76.13	54.67	89.10	70.59
DPN	80.93	48.60	85.29	81.07	92.97	77.54
GeoID (OURS)	**82.67**	**78.42**	**92.37**	**84.66**	**95.12**	**92.78**

Table 8: Known class accuracy and unknown class accuracy across different datasets. The known class ratio is set to 0.75 for fair comparison.