A Unified Knowledge Transfer Network for Generalized Category Discovery

Wenkai Shi[1,3],**Wenbin An**[1,3],**Feng Tian**[2,3*],**Yan Chen**[2],**Yaqiang Wu**[4],**Qianying Wang**[4],**Ping Chen**[5]

[1] School of Automation Science and Engineering, Xi'an Jiaotong University
[2] School of Computer Science and Technology, Xi'an Jiaotong University
[3] National Engineering Laboratory for Big Data Analytics
[4] Lenovo Research
[5] Department of Engineering, University of Massachusetts Boston
shiyibai778@gmail.com, {fengtian,chenyan}@mail.xjtu.edu.cn
wenbinan@stu.xjtu.edu.cn, {wuyqe, wangqya}@lenovo.com, ping.chen@umb.edu

Abstract

Generalized Category Discovery (GCD) aims to recognize both known and novel categories in an unlabeled dataset by leveraging another labeled dataset with only known categories. Without considering knowledge transfer from known to novel categories, current methods usually perform poorly on novel categories due to the lack of corresponding supervision. To mitigate this issue, we propose a unified **K**nowledge **T**ransfer **N**etwork (KTN), which solves two obstacles to knowledge transfer in GCD. First, the mixture of known and novel categories in unlabeled data makes it difficult to identify transfer candidates (i.e., samples with novel categories). For this, we propose an entropy-based method that leverages knowledge in the pre-trained classifier to differentiate known and novel categories without requiring extra data or parameters. Second, the lack of prior knowledge of novel categories presents challenges in quantifying semantic relationships between categories to decide the transfer weights. For this, we model different categories with prototypes and treat their similarities as transfer weights to measure the semantic similarities between categories. On the basis of two treatments, we transfer knowledge from known to novel categories by conducting pre-adjustment of logits and post-adjustment of labels for transfer candidates based on the transfer weights between different categories. With the weighted adjustment, KTN can generate more accurate pseudo-labels for unlabeled data, which helps to learn more discriminative features and boost model performance on novel categories. Extensive experiments show that our method outperforms state-of-the-art models on all evaluation metrics across multiple benchmark datasets. Furthermore, different from previous clustering-based methods that can only work offline with abundant data, KTN can be deployed online conveniently with faster inference speed. Code and data are available at https://github.com/yibai-shi/KTN.

Introduction

Although current machine learning methods have achieved impressive performance on many NLP tasks, they often fail to meet requirements in the real world. For instance, general text classification models trained on pre-defined categories cannot discover novel categories from unlabeled corpus. To cope with this limitation, many works in NLP (Lin, Xu, and Zhang 2020; Zhang et al. 2021; An et al. 2023a) and CV (Yu et al. 2022) were conducted, and Vaze et al. (2022) formalized the setting of these works as Generalized Category Discovery (GCD). GCD aims to identify both known and novel categories in unlabeled data, leveraging labeled data with only known categories, which can help models tackle unpredictable novel categories without extra labeling costs.

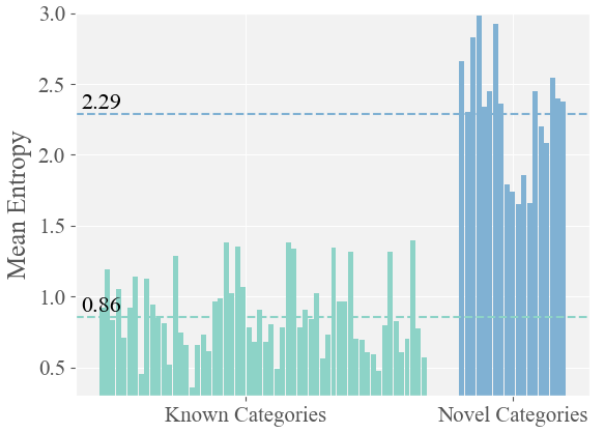


Figure 1: Prediction entropy (bar) of pre-trained classifier and corresponding mean value (dotted line) on BANKING dataset. Green ones are for known categories, and blue ones are for novel categories.

Existing GCD methods (Zhang et al. 2021; An et al. 2023b; Yu et al. 2022; Vaze et al. 2022) usually adopt a two-stage training strategy: pre-training on labeled data, then clustering unlabeled data with acquired knowledge in a semi-supervised manner. However, without transferring knowledge from known to novel categories, these methods usually perform poorly on novel categories due to the lack of corresponding supervision. Intuitively, although transferring knowledge from known to novel categories helps compensate for the supervision, this transfer will be hindered by two obstacles in GCD settings. The first is the **ambiguity of transfer candidates** (i.e., samples with novel categories). Considering the dilemma where known and novel categories

*Corresponding Authors.

are mingled in unlabeled data, the bias towards known categories brought by pre-training makes novel category samples easy to be confused with known category samples and difficult to distinguish. The ambiguity of transfer candidates causes redundant disturbance to the outputs of known category samples, inevitably obscuring decision boundaries within known categories. The second is the **uncertainty of transfer weights**. Since different categories have intrinsic semantic correlations (e.g., both dogs and foxes have thick fur; thus dog features can help discover foxes but not unrelated cars), it is necessary to determine the transfer weights to guide knowledge from known to novel categories with similar semantics. However, due to the lack of information such as ground-truth labels, it is difficult to directly measure the semantic similarities between categories to decide the transfer weights. The uncertainty of transfer weights leads to negative transfer to dissimilar novel categories, inevitably obscuring decision boundaries within novel categories.

To ensure an effective knowledge transfer, we propose a **K**nowledge **T**ransfer **N**etwork (KTN), which can sufficiently transfer knowledge from known to novel categories. For the ambiguity of transfer candidates, we propose **E**ntropy-based **S**oft **D**ifferentiation (ESD). As shown in Figure 1, the mean entropy of the pre-trained classifier's predictions for novel category samples (blue bar) is much higher than those for known category samples (green bar). On this basis, we compare the prediction entropy of unlabeled instances to differentiate known and novel categories within them. To alleviate the negative effect of differentiation noise brought by threshold-based hard differentiation, we map the prediction entropy to the probability that a sample belongs to novel categories in order to identify transfer candidates in a soft manner. For the uncertainty of transfer weights, we model different categories with corresponding prototypes in the embedding space and treat their similarities as transfer weights to measure semantic similarities between known and novel categories. After identifying the transfer candidates and transfer weights, we transfer knowledge from known to novel categories by adjusting the pseudo-label for each candidate on logits and label level, based on the transfer weights between different categories. With more accurate pseudo-labels, KTN can learn more discriminative features and boost model performance on novel categories.

Furthermore, different from previous clustering-based methods that can solely work offline with abundant data, KTN can be deployed online conveniently with faster inference speed.

Our main contributions can be summarized as follows:

- We first introduce the idea of explicit knowledge transfer from known to novel categories and thus propose an easy-to-deploy **K**nowledge **T**ransfer **N**etwork (KTN) for GCD, which can boost performance on novel categories through adjusted pseudo-labels.
- We propose an entropy-based method (ESD) to identify transfer candidates and model categories with prototypes to determine transfer weights, which avoids negative transfer to acquire better decision boundaries within known and novel categories.
- We conduct extensive experiments on multiple benchmark datasets to verify the effectiveness of our method.

Related Work

Generalized Category Discovery

Generalized Category Discovery (GCD) is formalized by (Vaze et al. 2022), which can be seen as a natural extension of NCD (Han, Vedaldi, and Zisserman 2019) in the open world. Different from NCD which has no category overlap between labeled and unlabeled data, GCD assumes that known and novel categories are mingled in unlabeled data, making models prone to overfit labeled known categories. To tackle GCD, Lin, Xu, and Zhang (2020); Zhang et al. (2021, 2022) utilized pair-wise similarity prediction, aligned pseudo-labels, and neighborhood relationships in the embedding space to cluster unlabeled data and discover novel categories. Vaze et al. (2022) introduced supervised and self-supervised contrastive learning to improve representation quality, and (An et al. 2023b) proposed a decoupled prototypical network for better representation learning. Although these methods can improve the performance on known categories, they still perform poorly on novel categories due to the lack of supervision, which requires effective knowledge transfer from known to novel categories.

Semi-Supervised Learning

Different from NCD or GCD, Semi-Supervised Learning (SSL) assumes that labeled data and unlabeled data are completely coincident in category distribution. Consistency Regularization is a widely adopted strategy in SSL, which enforces models to output consistent predictions for different augmentations of the same instance. For example, Laine and Aila (2016) aggregated multiple previous predictions as an augmentation of inputs, and Tarvainen and Valpola (2017) replaced it with the outputs of an EMA model. MixMatch (Berthelot et al. 2019b), ReMixMatch (Berthelot et al. 2019a), and FixMatch (Sohn et al. 2020) adopted different sampling techniques and data augmentations to further leverage the augmentation consistency. Although these methods cannot be directly applied to GCD because of category distribution misalignment between labeled and unlabeled data, they are enlightening to help models learn discriminative features with limited labeled data.

Method

Problem Formulation

Generalized Category Discovery (GCD) follows an open-world setting, which aims to recognize both known categories $\mathcal{Y}^k$ and novel categories $\mathcal{Y}^n$ ($\mathcal{Y}^k \cap \mathcal{Y}^n = \emptyset$) with a labeled dataset with only known categories $\mathcal{D}^l = \{(x_i, y_i)|y_i \in \mathcal{Y}^k\}$ and an unlabeled dataset containing all categories $\mathcal{D}^u = \{x_i|y_i \in \mathcal{Y}^k \cup \mathcal{Y}^n\}$. Finally, the model performance will be evaluated on the test set $\mathcal{D}^t = \{x_i|y_i \in \mathcal{Y}^k \cup \mathcal{Y}^n\}$.

Approach Overview

Figure 2 illustrates the overall architecture of our proposed Knowledge Transfer Network (KTN). During the first-stage

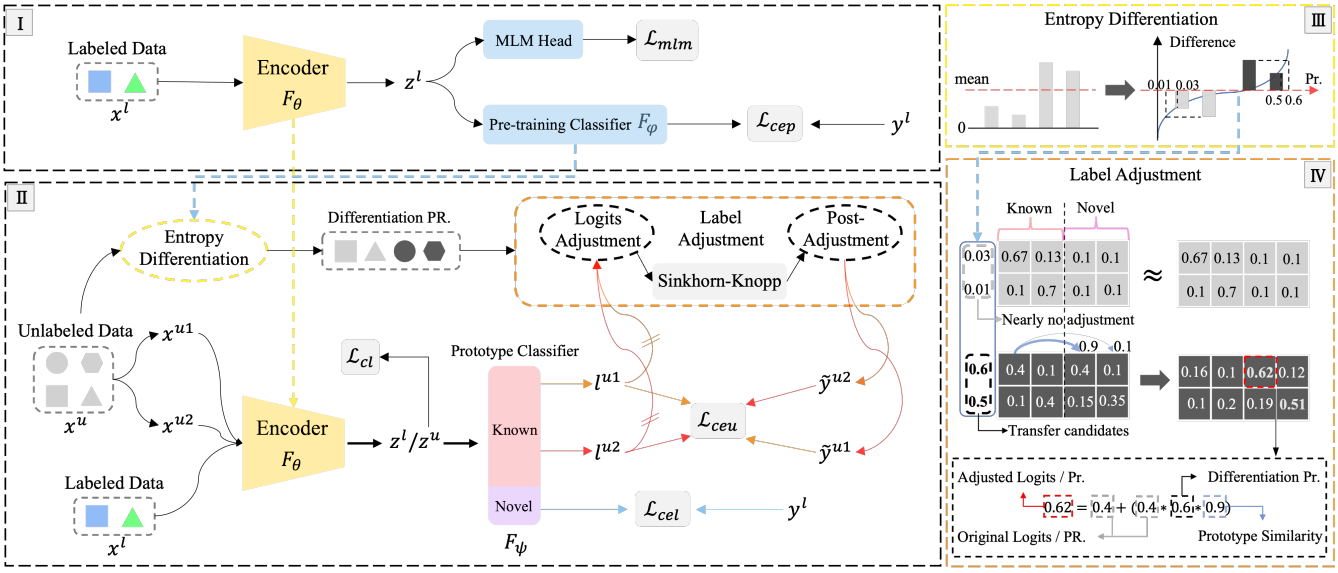


Figure 2: Overall architecture of Knowledge Transfer Network. x^l is the labeled instance, x^{u1} and x^{u2} are different views of the same unlabeled instance x^u. z, l, y, and $\tilde{y}$ refer to the feature, logits, ground-truth label, and pseudo-label.

pre-training shown in I, we utilize cross-entropy loss and masked language modeling loss to initialize the model. During the second-stage fine-tuning shown in II, we identify transfer candidates with entropy-based differentiation in III and obtain more accurate pseudo-labels with label adjustment in IV. The model is trained with unified cross-entropy loss on all training data.

Model Pre-training

We initialize the model with BERT encoder (Devlin et al. 2019) $F_\theta : \mathcal{X} \to \mathbb{R}^d$ and the pre-trained classifier $F_\varphi : \mathbb{R}^d \to \mathbb{R}^{K_1}$. Here d and $K_1 = |\mathcal{Y}^k|$ refer to the feature dimensions and the number of known categories, respectively. Following Zhang et al. (2022), we pre-train the model with cross-entropy loss on labeled data to acquire knowledge from known categories and masked language modeling loss on all data to learn domain-specific semantics:

$$\mathcal{L}_{pre} = \mathcal{L}_{cep}(\mathcal{D}^l) + \mathcal{L}_{mlm}(\mathcal{D}^l; \mathcal{D}^u) \tag{1}$$

where $\mathcal{D}^l$ and $\mathcal{D}^u$ are labeled and unlabeled training dataset, respectively.

Classifier Initialization

After pre-training, we first perform KMeans clustering on unlabeled data to get K cluster centroids as prototypes to represent corresponding categories (Snell, Swersky, and Zemel 2017), where $K = |\mathcal{Y}^k \cup \mathcal{Y}^n|$ is the number of total categories. Following An et al. (2023b), we decouple known and novel categories in prototypes and obtain prototype subsets corresponding to known and novel categories $\{\mu_i^k\}_{i=1}^{K_1}$ and $\{\mu_i^n\}_{i=1}^{K_2}$ respectively, where $K_2 = |\mathcal{Y}^n|$ is the number of novel categories. Then we fill the first K_1 dimensions of the prototype classifier with $\{\mu_i^k\}_{i=1}^{K_1}$ and fill remained K_2 dimensions with $\{\mu_i^n\}_{i=1}^{K_2}$. Finally, we $l2$-normalize the parameters of the unified prototype classifier $F_\psi : \mathbb{R}^d \to \mathbb{R}^K$ to complete initialization:

$$\psi = \left[\frac{\mu_1^k}{\|\mu_1^k\|}, \cdots, \frac{\mu_{K_1}^k}{\|\mu_{K_1}^k\|}, \frac{\mu_1^n}{\|\mu_1^n\|}, \cdots, \frac{\mu_{K_2}^n}{\|\mu_{K_2}^n\|} \right] \tag{2}$$

Notably, we presume prior knowledge of K following previous works (Zhang et al. 2021; An et al. 2023b) for a fair comparison.

Different from the utilization of decoupled prototypes in An et al. (2023b) which aims at direct model training in the embedding space and one-to-one knowledge transfer within known categories, we adopt the decoupling strategy so the prototype classifier can have a clear boundary between known and novel categories to achieve unambiguous knowledge transfer. Moreover, the classifier can distinguish samples in an online manner with faster inference speed compared with the offline clustering-based inference adopted by previous methods.

Entropy-based Soft Differentiation

Considering the pre-trained classifier is solely developed based on labeled known category data, it generally makes biased predictions, which are high confidence for known category instances but ambiguous for novel category instances. Since the high confidence of prediction represents low information entropy and vice versa, we compare prediction entropy to identify transfer candidates. Specifically, we first compute the prediction entropy of current mini-batch samples with the frozen pre-trained classifier:

$$e_i = -p_i \log p_i \tag{3}$$

where $p_i = \text{SoftMax}(F_\varphi(F_\theta(x_i)))$ is the posterior probability distribution over known categories of the i-th sample x_i,

e_i is the corresponding prediction entropy. Then, as shown in Figure 2 III, instead of setting a fixed threshold to differentiate known and novel categories in unlabeled data, we map the entropy difference between the samples and their mean value to the differentiation probability that a sample belongs to novel categories:

$$\omega_i = \frac{1}{1 + e^{-(d_i - \alpha * d_{\max})}} \tag{4}$$

where $d_i = e_i - \frac{1}{B}\sum_{j=1}^{B} e_i$ is the entropy difference of x_i, B is the current mini-batch size, $d_{\max}$ is the maximum difference, α is a hyperparameter to adjust the probability. Our proposed ESD retains the pre-trained classifier as the discriminator of transfer candidates without requiring any extra data or parameters. Moreover, compared with hard differentiation, ESD can alleviate the negative effect of differentiation noise among samples around the mean entropy.

Prototype-based Transfer Weights

Due to the lack of prior information such as ground-truth labels, intrinsic semantic correlations between categories cannot be directly quantified, which hinders effective knowledge transfer to novel categories with similar semantics. To this end, we propose to compute normalized cosine similarities between category prototypes to measure semantic similarities between known and novel categories:

$$P_{ij} = \frac{e^{cos(\mu_i^k, \mu_j^n)/\tau_1}}{\sum_{k=1}^{K_2} e^{cos(\mu_i^k, \mu_k^n)/\tau_1}} \tag{5}$$

where P_{ij} denotes the similarity between i-th known category prototype and j-th novel category prototype, τ_1 is a temperature hyperparameter. For the simplicity of subsequent representations, we zero-pad the semantic similarity matrix $\hat{P} \in \mathbb{R}_+^{K \times K}$ as below:

$$\hat{P} = \begin{bmatrix} \mathbf{0}_{K_1 \times K_1} & P \\ \mathbf{0}_{K_2 \times K_1} & \mathbf{0}_{K_2 \times K_2} \end{bmatrix} \tag{6}$$

Since higher semantic similarity means higher transferability, prototype-based transfer weights clarify target categories, avoiding negative transfer to semantically dissimilar novel categories. For simplicity, we solely compute the semantic similarity matrix once before the second-stage finetuning.

Label Adjustment

After identifying the transfer candidates and transfer weights, we adjust the pseudo-label assignment for each candidate on logits and label level. First, we extract sample feature $z_i = F_\theta(x_i)$ and obtain corresponding logits $l_i = F_\psi(z_i)$ by the prototype classifier. For each candidate, we select top-m values in logits belonging to known categories as our transfer outsets since their corresponding categories generally have the most similar semantics with the novel category to be discovered, which confuses discrimination of the classifier. We denote the index matrix $Q \in \mathbb{R}_+^{B \times K}$ of transfer outsets as:

$$Q_{ij} = \begin{cases} 1, & j \in argtop_m(\{l_{ik}\}_{k=1}^{K_1}) \\ 0, & otherwise \end{cases} \tag{7}$$

Dataset	$\lvert\mathcal{Y}^k\rvert$	$\lvert\mathcal{Y}^n\rvert$	$\lvert\mathcal{D}^l\rvert$	$\lvert\mathcal{D}^u\rvert$	$\lvert\mathcal{D}^t\rvert$
BANKING	58	19	673	8330	3080
CLINC	113	37	1344	16656	2250
HWU64	48	16	578	7134	1033

Table 1: Statistics of datasets. $|\mathcal{Y}^k|$, $|\mathcal{Y}^n|$, $|\mathcal{D}^l|$, $|\mathcal{D}^u|$ and $|\mathcal{D}^t|$ represent the number of known categories, novel categories, labeled data, unlabeled data, and test data.

where $argtop_m$ contains the index of top-m values. Second, as shown in Figure 2 IV, we reduce the selected values in l_i in proportion to the differentiation probability and compensate them to logits belonging to novel categories based on transfer weights. We denote the transfer process on logits level as:

$$\begin{aligned} \tilde{L} &= L - W(Q \otimes L) + W(Q \otimes L)\hat{P} \\ &= L + W(Q \otimes L)(\hat{P} - I) \end{aligned} \tag{8}$$

where $L = [l_1, \cdots, l_B]^\top \in \mathbb{R}_+^{B \times K}$ is the logtis matrix of current mini-batch samples, and $W = diag(w_1, \cdots, w_B) \in \mathbb{R}_+^{B \times B}$ is the differentiation probability matrix.

To avoid degenerate solutions, we transform pseudo-label assignment to an instance of optimal transport following Asano, Rupprecht, and Vedaldi (2019), which helps generate pseudo-labels online with adjusted logits:

$$\max_{Y \in \mathcal{Y}} \text{Tr}(Y\tilde{L}) + \epsilon H(Y) \tag{9}$$

where $Y \in {\mathbb{R}_+}^{B \times K}$ is the mapping matrix, $\epsilon H(Y) = -\epsilon \sum_{ij} Y_{ij} \log Y_{ij}$ is the weighted entropy term that enforces the equipartition of samples. The mapping matrix is constrained by the following transportation polytope $\mathcal{Y}$:

$$\mathcal{Y} = \left\{ Y \in \mathbb{R}_+^{B \times K} | Y^\top \mathbf{1}_B = \frac{1}{K}\mathbf{1}_B, Y\mathbf{1}_K = \frac{1}{B}\mathbf{1}_K \right\} \tag{10}$$

where $\mathbf{1}_K$ denotes the vector of ones in dimension K. We utilize the Sinkhorn-Knopp algorithm (more details can be referred to Cuturi (2013)) to obtain the optimal mapping Y^* as our pseudo-labels.

In order to further eliminate the ambiguity between known and novel categories for pseudo-labels, apart from pre-adjustment on logits level, we also perform post-adjustment on label level. Specifically, we adopt the same adjustment strategy as Eq. 8 to conduct post-adjustment:

$$\tilde{Y} = Y^* + W(Q \otimes Y^*)(\hat{P} - I) \tag{11}$$

Then, the pseudo-label whose largest novel category probability falls above a pre-defined threshold γ is converted to a one-hot label, which helps make high-confidence predictions for transfer candidates.

Our proposed label adjustment strategy effectively transfers knowledge from known to novel categories, which can help KTN 1) overcome bias to known categories with selective logits/probabilities reduction and 2) avoid discrimination confusion within novel categories with weighted logits/probabilities compensation.

Method	BANKING			HWU64			CLINC		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
DeepCluster	13.95	13.94	13.99	16.24	17.04	13.84	26.92	27.34	25.67
DCN	17.85	18.94	14.35	20.37	21.63	16.59	29.64	30.00	28.45
DEC	19.30	20.36	15.84	21.63	23.47	16.11	19.99	20.18	19.40
GloVe-KM	29.18	29.11	29.39	35.42	35.08	36.44	51.64	51.74	51.50
SAE-KM	38.05	38.29	37.27	42.26	44.53	35.45	46.59	47.35	44.24
Semi-DC	50.73	53.37	42.63	52.91	54.41	48.42	74.52	75.60	71.34
CDAC+	53.09	55.42	46.01	56.82	58.14	52.86	69.75	70.08	68.77
DTC	56.56	59.98	46.10	62.56	64.94	55.42	76.42	82.34	58.95
Semi-KM	66.23	73.62	43.68	71.71	82.32	39.69	81.42	89.03	59.01
DAC	63.63	69.60	45.44	72.09	77.55	55.64	84.42	89.10	70.59
GCD	67.55	75.16	44.34	73.64	78.58	58.82	79.26	89.64	48.66
DPN	72.96	80.93	48.60	74.29	78.84	60.64	89.06	92.97	77.54
KTN (Ours)	**77.93**	**83.85**	**60.18** (+11.58)	**81.32**	**84.41**	**72.07** (+11.43)	**91.28**	**94.43**	**81.84** (+4.3)

Table 2: Average results (%) over 3 runs on test sets.

Overall Training Objective

For unlabeled instance x_i^u, we retain the original input as first view x_i^{u1} and generate second view x_i^{u2} with data augmentation. Their corresponding pseudo-labels $\tilde{y}_i^{u1}$ and $\tilde{y}_i^{u2}$ are obtained by Label Adjustment. To alleviate overfitting and improve prediction robustness, we enforce consistency between two views from feature level with infoNCE loss:

$$\mathcal{L}_{cl} = -\sum_{i=1}^{B^u} \log \frac{\exp(z_i^{u1} \cdot z_i^{u2}/\tau_2)}{\sum_j \mathbb{1}_{[j \neq i]} \exp(z_i^{u1} \cdot z_j^{u2}/\tau_2)} \quad (12)$$

and consistency from label level with swapped prediction:

$$\mathcal{L}_{ceu} = -\sum_{i=1}^{B^u} \tilde{y}_i^{u1} \cdot \log(p_i^{u2}) - \sum_{i=1}^{B^u} \tilde{y}_i^{u2} \cdot \log(p_i^{u1}) \quad (13)$$

where z_i^{u1} and z_i^{u2} are embeddings obtained by F_θ, p_i^{u1} and p_i^{u2} are predictions obtained by F_ϕ, τ_2 is a temperature hyperparameter, B^u is the mini-batch of unlabeled instances. To avoid catastrophic forgetting of known categories, we also add cross-entropy loss on labeled data:

$$\mathcal{L}_{cel} = -\sum_{i=1}^{B^l} y_i^l \cdot \log(p_i^l) \quad (14)$$

where B^l is the mini-batch of labeled instances, y_i^l is the ground-truth label of x_i^l. Overall, the training objective of KTN is formulated as follows:

$$\mathcal{L}_{KTN} = \lambda_1 \mathcal{L}_{ceu} + (1 - \lambda_1)\mathcal{L}_{cel} + \lambda_2 \mathcal{L}_{cl} \quad (15)$$

where λ_1 and λ_2 control the weights of losses.

Experiments

Datasets

We evaluate our method on three benchmark datasets. **BANKING** is a fine-grained intent classification dataset released by Casanueva et al. (2020). **CLINC** is a multi-domain intent classification dataset released by Larson et al. (2019). **HWU64** is a personal assistant query classification dataset released by Liu et al. (2021). More details of these datasets are summarized in Table 1.

Comparison Methods

We compare our method with various baselines and state-of-the-art methods.

Unsupervised Methods. GloVe-KM: KMeans with GloVe embeddings (Pennington, Socher, and Manning 2014); SAE-KM: KMeans with embeddings learned by stacked auto-encoder; DEC: Deep Embedded Clustering (Xie, Girshick, and Farhadi 2016); DCN: Deep Clustering Network (Yang et al. 2017); DeepCluster: Deep Clustering (Caron et al. 2018).

Semi-supervised Methods. Semi-KM: KMeans with BERT pretrained on labeled data; Semi-DeepCluster: Deep Clustering pretrained on labeled data; DTC: Deep Transfer Clustering (Han, Vedaldi, and Zisserman 2019); CDAC+: Constrained Adaptive Clustering (Lin, Xu, and Zhang 2020); DAC: Deep Aligned Clustering (Zhang et al. 2021); GCD: Label Assignment with Semi-supervised KMeans (Vaze et al. 2022); DPN: Decoupled Prototypical Network (An et al. 2023b).

Evaluation Metrics

We adopt three metrics to evaluate the model performance: classification accuracy for known category instances (Known), clustering accuracy for novel category instances (Novel), and the overall accuracy across the test set (All).

Implementation Details

We use the pre-trained BERT model (bert-base-uncased) as our backbone and AdamW optimizer with 0.01 weight decay. During pre-training, we adopt an early-stopping strategy with a patience of 20 epochs. Furthermore, we retain the pre-trained encoder and classifier as our discriminator of

Method	All	Known	Novel
KTN	77.93	83.85	60.18
w/o infoNCE loss $\mathcal{L}_{cl}$	77.48	83.44	59.96
w/o swapped prediction	77.29	82.53	61.57
w/o CE loss $\mathcal{L}_{cel}$	72.64	75.03	65.48
w/o classifier initialization	72.21	82.12	42.47
w/o label adjustment	71.98	84.07	35.71

Table 3: Ablation results (%) of different model variants.

logits-	**label-**	**ESD**	**SW**	All	Known	Novel
✗	✓	✓	✓	76.68	84.86	52.12
✓	✗	✓	✓	74.44	84.21	45.13
✓	✓	✗	✗	73.21	77.16	61.32
✓	✓	✗	✓	74.06	77.04	65.12
✓	✓	✓	✗	76.70	83.21	57.16
✓	✓	✓†	✓	77.05	82.34	61.18
✓	✓	✓	✓	77.93	83.85	60.18

Table 4: Ablation results (%) of the label adjustment strategy. † means hard differentiation by a fixed threshold.

known and novel categories. During the second-stage fine-tuning, we set $\alpha = 1$ in Eq. 4 to determine the differentiation probability, $\tau_1 = 1$ in Eq. 5 to determine semantic weights between different categories, $m = 1$ in Eq. 7 as the number of transfer outsets, $\gamma = 0.5$ to filter high-confidence instances with novel categories, $\tau_2 = 0.07$ in Eq. 12 and loss weights $\{\lambda_1, \lambda_2\} = \{0.7, 0.01\}$ in Eq. 15 to balance different losses. Moreover, we adopt random token replacement (Zhang et al. 2022) as data augmentation for second view generation. For other general hyperparameters, the learning rate is set to $5e^{-5}$, training epochs are set to 80, and the batch size of labeled and unlabeled instances is set to 128 for all datasets equally. For the implementation of the Sinkhorn-Knopp algorithm, we inherit all parameters from Caron et al. (2020) directly.

Results and Discussion

Main Results

The main results are shown in Table 2. Our proposed KTN outperforms previous methods on all evaluation metrics and benchmark datasets by a large margin. The **4.97%**, **7.03%**, and **2.22%** improvement on overall accuracy reflect the effectiveness of our method, which strikes a balance between known and novel categories. Specifically, KTN outperforms the SOTA model by **2.92%**, **5.57%**, and **1.46%** on accuracy for known category instances. Apart from the cross-entropy loss on labeled data, this improvement can also be attributed to KTN identifying transfer candidates, which reduces redundant disturbance to instances with known categories and avoids the ambiguity of decision boundaries within known categories. More importantly, KTN achieves **11.58%**, **11.43%**, and **4.3%** improvement on accuracy for novel category instances compared with the SOTA model, which fully demonstrates that our label adjustment strategy based on transfer candidates and semantic weights effectively transfer knowledge from known to novel categories to compensate for the supervision for novel categories.

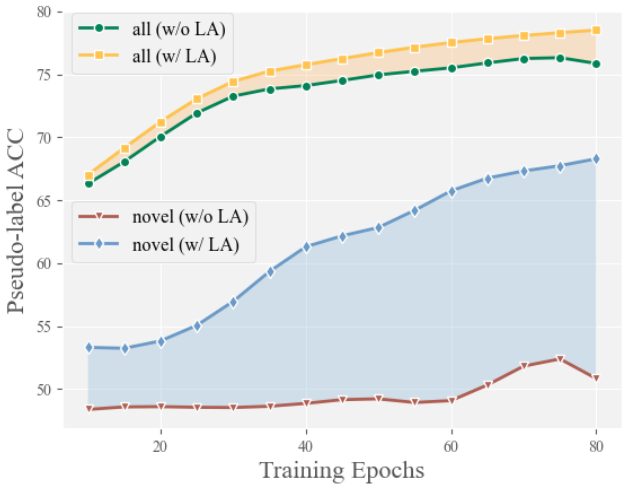


Figure 3: Pseudo-label accuracy (%) w/ and w/o label adjustment on BANKING dataset.

Ablation Study

The performance of different model variants on BANKING is shown in Table 3. All components affect model performance more or less. Specifically, 1) removing infoNCE loss $\mathcal{L}_{cl}$ (Eq. 12) or swapped prediction in Eq. 13 slightly hinders model learning discriminative features; 2) removing CE loss on labeled data (Eq. 14) leads to **8.82 %** decline on accuracy for known category instances due to catastrophic forgetting; 3) removing classifier initialization (Eq. 2) degrades performance on novel categories (**17.71 %**) owing to the unclear boundary in classifier hindering knowledge transfer. 4) removing label adjustment strategy (Eq. 8 and Eq. 11) seriously impairs knowledge transfer from known to novel categories, resulting in **24.47%** decline on accuracy for novel category instances;

Analysis of Label Adjustment

Since Label Adjustment helps knowledge transfer, we further investigate the effect of different components on Banking: pre-adjustment of logits, post-adjustment of pseudo-labels, transfer candidates based on ESD, and prototype-based transfer weights. The ablation results shown in Table 4 indicate that all components contribute to the overall accuracy. In detail, 1) removing either adjustment on logits or label level seriously degrades the clustering accuracy for novel categories, which demonstrates the effectiveness of adjusted pseudo-labels in knowledge transfer; 2) treating all unlabeled instances as transfer candidates instead of ESD impairs performance on known categories due to ambiguous decision boundaries within known categories brought by redundant disturbance; 3) replacing semantic similarities

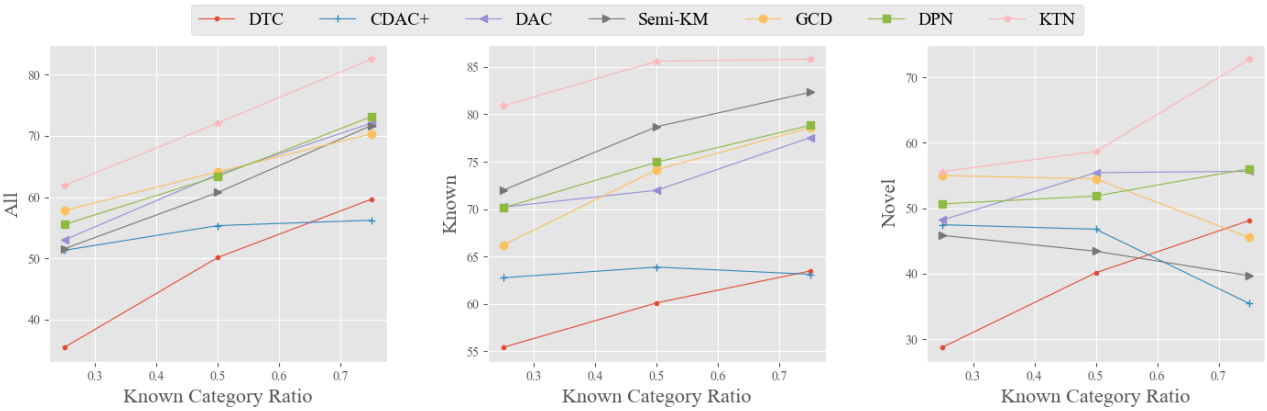
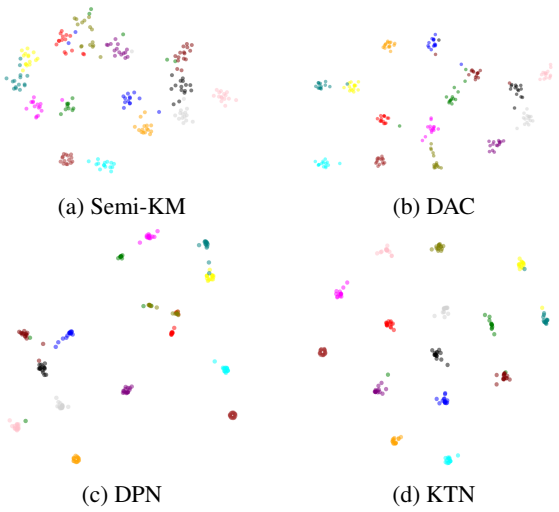


Figure 4: Influence of known category ratio on HWU64 dataset.



(a) Semi-KM (b) DAC

(c) DPN (d) KTN

Figure 5: The t-SNE visualizations of embeddings.

with uniform distribution as transfer weights obscures decision boundaries within novel categories, impairing performance on novel categories; 4) utilizing mean entropy as a fixed threshold to differentiate unlabeled data degrades performance on known categories owing to the negative transfer to noisy instances with known categories.

Quality of Pseudo-labels

Figure 3 intuitively reflects the change of pseudo-label quality on BANKING. In detail, with our proposed adjustment strategy, the overall accuracy of pseudo-labels is continuously improved with the training progress (orange area), especially the accuracy on novel categories (blue area). This experimental phenomenon illustrates that KTN effectively transfers knowledge from known to novel categories to compensate for the supervision of novel categories.

Effect of Known Category Ratio

To investigate the influence of the known category ratio on model performance, we vary it in the range of 0.25, 0.50, and 0.75. As shown in Figure 4, our method achieves comparable or best performance under different settings on all evaluation metrics, which fully demonstrates the effectiveness and robustness of KTN.

Method	BANKING	HWU	CLINC
DAC	17.68	4.34	15.92
KTN	3.12	0.39	1.12

Table 5: Efficiency (sec.) comparison of model inference.

Feature Visualization

As shown in Figure 5, we use t-SNE to visualize embedding learned by different methods on CLINC dataset (15 fine-grained categories under the same coarse-grained category). The visualization results clearly show more separable and compact feature distributions compared with other methods, which fully demonstrates the effectiveness of KTN.

Comparison of Inference Efficiency

Table 5 illustrates the time cost of inference with different methods across benchmark datasets. Different from the clustering-based inference that must traverse abundant data offline multiple times, KTN can complete online inference for any number of samples by solely one forward propagation, which significantly reduces the average inference time.

Conclusion

In this paper, we first introduce the idea of explicit knowledge transfer from known to novel categories and propose a **K**nowledge **T**ransfer **N**etwork (KTN) for GCD. Furthermore, we creatively use the prediction entropy of the pre-trained classifier (ESD) to identify transfer candidates and determine prototype-based transfer weights to measure semantic similarities between categories, which help adjust pseudo-labels to learn more discriminative features and boost performance on novel categories. Experimental results on multiple datasets illustrate that KTN outperforms SOTA methods by a large margin.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2022ZD0117102), National Natural Science Foundation of China (62177038, 62293551, 62277042, 62137002, 61937001, 62377038). Project of China Knowledge Centre for Engineering Science and Technology, "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project.

References

An, W.; Tian, F.; Chen, P.; Zheng, Q.; and Ding, W. 2023a. New User Intent Discovery with Robust Pseudo Label Training and Source Domain Joint-training. *IEEE Intelligent Systems*.

An, W.; Tian, F.; Zheng, Q.; Ding, W.; Wang, Q.; and Chen, P. 2023b. Generalized category discovery with decoupled prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12527–12535.

Asano, Y.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*.

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

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

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Casanueva, I.; Temčinas, T.; Gerz, D.; Henderson, M.; and Vulić, I. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 38–45.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

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

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Larson, S.; Mahendran, A.; Peper, J. J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J. K.; Leach, K.; Laurenzano, M. A.; Tang, L.; 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)*, 1311–1316.

Lin, T.-E.; Xu, H.; and Zhang, H. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8360–8367.

Liu, X.; Eshghi, A.; Swietojanski, P.; and Rieser, V. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, 165–183. Springer.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

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

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

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7492–7501.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3861–3870. PMLR.

Yu, Q.; Ikami, D.; Irie, G.; and Aizawa, K. 2022. Self-labeling framework for novel category discovery over domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3161–3169.

Zhang, H.; Xu, H.; Lin, T.-E.; and Lyu, R. 2021. Discovering new intents with deep aligned clustering. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 35, 14365–14373.

Zhang, Y.; Zhang, H.; Zhan, L.-M.; Wu, X.-M.; and Lam, A. 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)*, 256–269.