ProKD: An Unsupervised Prototypical Knowledge Distillation Network for Zero-Resource Cross-Lingual Named Entity Recognition

Ling Ge[1], Chunming Hu[1,2,*], Guanghui Ma[1], Hong Zhang[3], Jihong Liu[4]

[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] College of Software, Beihang University, Beijing, China
[3]National Computer Network Emergency Response Technical Team / Coordination Center of China, Beijing, China
[4]School of Mechanical Engineering and Automation, Beihang University, Beijing, China
{geling, hucm, maguanghui, ryukeiko}@buaa.edu.cn, zhangh@isc.org.cn

Abstract

For named entity recognition (NER) in zero-resource languages, utilizing knowledge distillation methods to transfer language-independent knowledge from the rich-resource source languages to zero-resource languages is an effective means. Typically, these approaches adopt a teacher-student architecture, where the teacher network is trained in the source language, and the student network seeks to learn knowledge from the teacher network and is expected to perform well in the target language. Despite the impressive performance achieved by these methods, we argue that they have two limitations. Firstly, the teacher network fails to effectively learn language-independent knowledge shared across languages due to the differences in the feature distribution between the source and target languages. Secondly, the student network acquires all of its knowledge from the teacher network and ignores the learning of target language-specific knowledge. Undesirably, these limitations would hinder the model's performance in the target language. This paper proposes an unsupervised prototype knowledge distillation network (ProKD) to address these issues. Specifically, ProKD presents a contrastive learning-based prototype alignment method to achieve class feature alignment by adjusting the distance among prototypes in the source and target languages, boosting the teacher network's capacity to acquire language-independent knowledge. In addition, ProKD introduces a prototypical self-training method to learn the intrinsic structure of the language by retraining the student network on the target data using samples' distance information from prototypes, thereby enhancing the student network's ability to acquire language-specific knowledge. Extensive experiments on three benchmark cross-lingual NER datasets demonstrate the effectiveness of our approach.

Introduction

Named Entity Recognition (NER) is a fundamental sub-task of information extraction that aims to locate and classify text spans into predefined entity classes such as locations, organizations, etc (Ma et al. 2022b). It is often employed as an essential component for tasks such as question answering (Cao et al. 2022) and coreference resolution(Ma et al. 2022a). Despite the impressive performance recently achieved by deep learning-based NER methods, these supervised methods are limited to a few languages with rich entity labels, such as English, due to the reasonably large amount of human-annotated training data required. In contrast, the total number of languages currently in use worldwide is about $7,000$ [1], the majority of which contain limited or no labeled data, constraining the application of existing methods to these languages (Wu et al. 2020c,b). Hence, cross-lingual transfer learning is gaining increasing attention from researchers, which can leverage knowledge from high-resource (source) languages(e.g., English) with abundant entity labels to overcome the data scarcity problem of the low- (zero-) resource (target) languages (Liu et al. 2021). In particular, this paper focuses on the zero-resource scenario, where there is no labeled data in the target language.

To improve the performance of zero-resource cross-lingual NER, researchers have conducted intensive research and proposed various approaches(Jain et al. 2019; Wu et al. 2020c; Pfeiffer et al. 2020). Among these, the knowledge distillation-based approaches (Chen et al. 2021; Wu et al. 2020b,a) have recently shown encouraging results.These approaches typically train a teacher NER network using source language data and then leverage the soft pseudo-labels produced by the teacher network for the target language data to train the student NER network. In this way, the student network is expected to learn the language-independent knowledge from the teacher network and perform well on unlabeled target data (Hinton, Vinyals, and Dean 2015).

While significant progress has been achieved by knowledge distillation-based approaches for cross-lingual NER, we argue that these approaches still have two limitations. First, knowledge distillation relies heavily on the shared language-independent knowledge acquired by the teacher network across languages. As is known, there are differences in the feature distribution between the source and target languages, existing techniques employ only the source language for teacher network training. As a result, the teacher network tends to learn source-language-specific knowledge and cannot effectively grasp shared language-independent knowledge. Second, under the knowledge distillation learning mechanism, the student network aims to match the pseudo-soft labels generated by the teacher network for the

*Corresponding author

[1]https://www.ethnologue.com/guides/how-many-languages

target language. Consequently, the student network acquires all of its knowledge from the teacher network and ignores the acquisition of target language-specific knowledge. Undesirably, these two constraints would hinder the model's performance in the target language.

In this paper, we propose an unsupervised **Pro**totypical **K**nowledge **D**istillation network (**ProKD**), which employs contrastive learning-based prototype alignment and prototypical self-training to address the two above limitations, respectively. Specifically, we rely on performing class-level alignment between the source and target languages in semantic space to enhance the teacher network's capacity for capturing language-independent knowledge. We argue that class-level alignment can bridge the gap in the feature distribution and force the teacher network better to learn the shared semantics of entity classes across languages (Van Nguyen et al. 2021; Xu et al. 2022). To do this, we choose prototypes (Snell et al. 2017), i.e., the class-wise feature centroids, rather than samples, for class-level alignment because prototypes are robust to outliers and friendly to class imbalance tasks (Qiu et al. 2021; Zhang et al. 2021). In order to pull the prototypes of the same class closer and push the prototypes of different classes far away across languages, we leverage classical contrastive learning (Chen et al. 2020) to adjust the distance among class prototypes. Thus the class-level representation alignment between the source and target languages is achieved.

Furthermore, we present a prototypical self-training method to enhance the student network's ability to acquire the target language-specific knowledge. In particular, we establish pseudo-hard labels for unlabeled target samples based on their softmax-valued relative distances, i.e., prototype probability, to all prototypes and then retrain the network using these pseudo-labels. Since the prototypes accurately represent the clustering distribution underlying the data, the prototypical self-training enables the student network to learn the intrinsic structure of the target language (Zhang et al. 2021), thus revealing language-specific knowledge, such as the token's label preference. In addition, while calculating the pseudo-hard labels, the class distribution probabilities generated by the teacher network are incorporated into the prototype probabilities to improve the quality of the pseudo-hard labels and facilitate self-training.

Summarily, we make four contributions: (1) We propose a ProKD model for zero-resource cross-lingual NER task, which can improve the model's generalization to the target language . (2) We propose a contrastive learning-based prototype alignment method to enhance the teacher network's ability to acquire language-independent knowledge. (3) We propose a prototypical self-training method to enhance the student network's ability to acquire target language-specific knowledge. (4) Experimental results on six target languages validate the effectiveness of our approach.

Related Works

Cross-lingual NER

Current research on cross-lingual NER with zero resources falls into three main branches. The translation-based methods rely on machine translation and label projection (Xie et al. 2018a; Jain et al. 2019) to construct pseudo-training data for the target language, all of which involve high human costs and introduce label noise. The direct transfer-based methods resort to training a NER model with the source language and directly transferring it to the target language (Wu and Dredze 2019; Wu et al. 2020c; Pfeiffer et al. 2020). These approaches fail to exploit information from the unlabelled target language, resulting in non-optimal cross-lingual performance. The knowledge distillation-based methods encourage the student network to learn language-independent knowledge from the teacher network. Specifically, Wu et al. (2020a) distills knowledge directly from multi-source languages. AdvPicker(Chen et al. 2021) leverages adversarial learning to select target data to alleviate the overfitting of the model to source data. We argue that the above approaches fail to effectively learn shared language-independent knowledge and ignore the acquisition of target language-specific knowledge.

Knowledge Distillation

Knowledge distillation enables knowledge transfer from the teacher network to the student network (Hinton, Vinyals, and Dean 2015), where the student network is optimized by fitting the soft labels generated by the trained teacher network. Since the soft targets have a high entropy value, they provide more information per training case than the hard targets (Hinton, Vinyals, and Dean 2015), the student network can learn from the teacher network and perform well on unlabeled data. Knowledge distillation achieves significant results in various tasks such as model compression(Liu et al. 2022), image classification (Hinton, Vinyals, and Dean 2015), dialogue generation (Peng et al. 2019), machine translation (Weng et al. 2020), etc. In this paper, we choose knowledge distillation as the basic framework of our proposed approach for zero-resource cross-lingual NER.

Methodology

The NER task is modeled as a sequence labeling problem in this paper, i.e., given a sentence $X = \{x_0, \cdots, x_i, \cdots, x_L\}$, the NER model is expected to produce a label sequence $Y = \{y_0, \cdots, y_i, \cdots, y_L\}$, where y_i denotes the entity class corresponding to token x_i. Following previous works' setting(Wu et al. 2020a,b), given a labeled dataset source language dataset, $\{(X^s_m, Y^s_m)\}^{n_s}_{m=1} \sim \mathcal{D}_s$, and an unlabeled target language dataset, $\{(X^t_m)\}^{n_t}_{m=1} \sim \mathcal{D}_t$, the zero-resource cross-lingual NER aims to train a model with the above two datasets and expects the model to obtain good performance on target language data.

Overall Architecture

In this section, we describe the proposed approach, ProKD, for cross-lingual NER with zero resource, whose architecture is shown in Fig 1 and Fig 2. The core of ProKD is a knowledge distillation framework that includes a teacher network and a student network. In more detail, the teacher network employs a prototype class alignment method based on contrastive learning, which enhances its ability to acquire

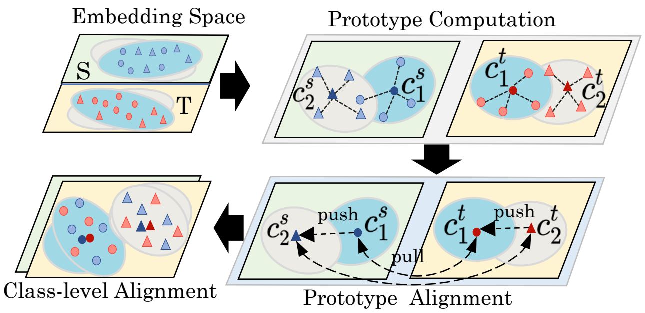


Figure 1: We achieve class-level feature alignment on the teacher network with a contrastive learning-based prototype alignment approach. The "S" and "T" denote the source and target languages, respectively.

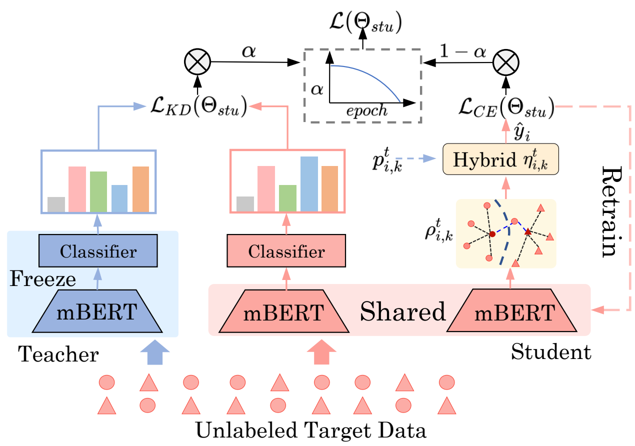


Figure 2: The student network can benefit from two aspects: the knowledge distillation and the self-training.

language-independent knowledge. The student network utilizes a prototypical self-training approach combined with the class distribution probability of the teacher network, which enhances its ability to learn language-specific knowledge.

Zero-resource Cross-Lingual NER via Knowledge Distillation

The knowledge distillation-based methods for zero-resource cross-lingual NER typically follow a two-stage training pipeline. First, the teacher network is trained with labeled source data, and then language-independent knowledge is distilled to the student network.

Given a sequence $X_m^s = \{x_0^s, \cdots, x_i^s, \cdots, x_L^s\}$ from source language data, the encoder f_θ of teacher network can map it into the hidden space and output the representations $H_m^s = \{h_0^s, \cdots, h_i^s, \cdots, h_L^s\}$. Following the previous works (Wu et al. 2020a,b), we adopt multilingual BERT(short for mBERT) (Devlin et al. 2019) as the feature encoder. Then we leverage a classifier with a softmax function to obtain the output p_i^s for each token x_i^s, and the cross entropy loss for the teacher network can be formulated as:

$$\mathcal{L}_{CE}(\Theta_{tea}) = -\frac{1}{n_s L} \sum_{(x^s, y^s) \in D_s} \sum_{i=0}^{L} y_i^s log(p_i^s) \quad (1)$$

where Θ_{tea} is the parameters of the teacher network to be optimized, n_s is the number of the sentences in dataset D_s, and y_i^s represents the golden label of token x_i^s.

Benefiting from the shared feature space of pre-trained mBERT and task knowledge from the labeled source data, we can directly utilize the teacher network to infer the class probabilities p_i^t of each token in a sequence X_m^t from unlabeled dataset D_t^x. Then the student network, consisting of a feature encoder mBERT and a classifier with a softmax function, is trained using these class probabilities as "soft targets" on the unlabeled dataset. To approximate the probabilities p_i^t , the training objective for the student network can be formulated as:

$$\mathcal{L}_{KD}(\Theta_{stu}) = \frac{1}{L} \sum_{x \in D_t} \sum_{i=1}^{L} (p_i^t - q_i^t)^2 \quad (2)$$

where p_i^t and q_i^t denote the probability distribution produced by the teacher and the student network for x_i^t, respectively. And here, following previous works (Yang et al. 2020; Wu et al. 2020a), we use the MSE loss to measure the prediction discrepancy of the two networks.

Prototypical Class-wise Alignment

Here, we present our method, prototypical class-wise alignment, to boost the teacher network's capacity to acquire language-independent knowledge.

Due to the absence of annotations on target language data, the class-wise alignment between the source and target languages is not trivial. To address this, as shown in Fig 1, we first calculate target class prototypes by class distribution probabilities produced by the teacher network in target data, and then leverage the prototype alignment between the two above languages to achieve class-wise alignment. We use prototype alignment rather than sample alignment since the prototype is robust to outliers (Zhang et al. 2021), and it can alleviate the negative impact of the noise (Xie et al. 2018b) introduced by the teacher network for the target data. Additionally, the prototype treats all classes equally (Zhang et al. 2021), which is crucial for the NER task, as non-entity type samples constitute the bulk of the overall samples.

To be specific, for the source language, we first obtain the token representation h_i^s of each token x_i^s using mBERT, and then, with the help of the golden labels, we directly compute the average representation of token samples with same label and treat it as the class prototype:

$$C_k^s = \frac{1}{n_k^s} \sum_{(X^s, Y^s) \in D_s} \sum_{i=0}^{L} \mathbb{I}[y_i^s = k] h_i^s \quad (3)$$

where k denotes an entity class label, $\mathbb{I}$ is an indicator function, and n_k^s represents the number of samples belonging to class k in the source language.

For the target language, we utilize the same method to obtain the representation h_i^t of each target token x_i^t. Since the target data is unlabeled, to alleviate the uncertainty of class prototype computation, we use the output of the teacher classifier to estimate the probabilities for the current token belonging to each class. Regarding these probabilities as weight, we aggregate representations of all target tokens to derive the target class prototype, which can be expressed as:

$$C_k^t = \frac{\sum_{X \in D_t} \sum_{i=0}^{L} p_{i,k}^t * h_i^t}{\sum_{X \in D_t} \sum_{i=0}^{L} p_{i,k}} \tag{4}$$

where $p_{i,k}^t$ represents the probability that the token x_i belongs to the class k.

The class prototype calculation involves all the samples, leading to high computing costs. To reduce the computation complexity while ensuring the stability of updates, we use the moving average method (Xie et al. 2018b) to update the source and target prototypes :

$$C_{k,cur}^{s(t)} = \lambda * C_{k,cur}^{s(t)} + (1-\lambda) * C_{k,cur-1}^{s(t)} \tag{5}$$

where $\lambda \in (0,1)$ is the moving average coefficient, cur denotes the current moment and $cur-1$ indicates the previous moment. In practical implementation, the source prototypes are updated once per epoch, while the target prototypes are updated once per batch.

After obtaining all class prototypes, we leverage classical contrastive learning to adjust the distance among prototypes in the feature space for class-wise alignment. For prototypes from source and target data with the same class, we regard one as an anchor (e.g., C_i^s) and the other as the positive sample of the anchor (e.g., C_i^t), while the rest of the prototypes are considered as negative samples (marked as $C_{i,neg}^s$). Then the class alignment loss is presented as:

$$\mathcal{L}_{CA}(\Theta_{tea}) = -\log \sum_{i=1}^{num} \frac{\exp(z_i^s \cdot z_i^t/\tau_1)}{\sum_{neg}\exp(z_i^s \cdot z_{i,neg}^s/\tau_1) + \sum_{neg}\exp(z_i^t \cdot z_{i,neg}^t/\tau_1)} \tag{6}$$

where z_i^s , z_i^t, $z_{i,neg}^s$ and $z_{i,neg}^t$ are l_2 regularization of C_i^s , C_i^t, $C_{i,neg}^s$, $C_{i,neg}^t$, respectively, the $C_{i,neg}^t$ denotes the negative samples of C_i^t, τ_1 is a temperature parameter, and num is the number of entity classes .

In this way, we can pull in source and target prototypes of the same class and push away source and target prototypes of different classes. Finally, we obtain the total loss $\mathcal{L}(\Theta_{tea})$ for the teacher network, consisting of the cross-entropy loss and the class alignment loss:

$$\mathcal{L}(\Theta_{tea}) = \mathcal{L}_{CE}(\Theta_{tea}) + \mathcal{L}_{CA}(\Theta_{tea}) \tag{7}$$

Prototypical Self-trainning

Here, we present our approach prototypical self-training with the unlabeled target language data, to boost the student network's ability to learn language-specific knowledge.

Specifically, we rely on prototype learning to iteratively generate hard pseudo labels for unlabelled target language samples and leverage these hard labels to conduct self-training on the target data. This is because the prototypes can perceive the underlying clustering distribution of the data, fundamentally reflecting the internal structure of the data and the intrinsic differences across the data (Zhang et al. 2021), which facilitates the learning of language-specific knowledge, such as the label preference of a token.

To acquire the target class prototypes, we first obtain the hidden representations and prediction probabilities through the student network, respectively, and then leverage the exact prototype computation and updating equation (Equation 4 and Equation 5) as the teacher network to obtain the class prototypes C^t. Afterwards, a class probability distribution ρ_i based on prototypes is calculated by leveraging the sample's feature distance w.r.t the class prototypes:

$$\rho_{i,k}^t = \frac{exp(-\|h_i^t - C_k^t\|/\tau_2)}{\sum_{k'} exp(-\|h_i^t - C_{k'}^t\|/\tau_2)} \tag{8}$$

where τ_2 is the softmax temperature, and $\rho_{i,k}^t$ represents the softmax probability of sample x_i belonging to the kth class. As observed, if a feature representation h_i^t is far from the prototype C_k^t, the probability of this feature for class k would be very low. We convert ρ_i^t into a hard pseudo-label $\hat{y}_i^t$ based on the following formula:

$$\hat{y}_i^t = \xi(\rho_i^t) \tag{9}$$

where ξ denotes the conversion function.

Intuitively, we can use these pseudo-hard labels for self-training. However, one natural question then arises. Prototypical self-training is essentially cluster-based representation learning and will inevitably introduce incorrect label in pseudo labeling. For instance, when a sample is far from the prototype to which it belongs, the student network may mislabel this sample(Snell et al. 2017). To alleviate this issue, we fuse the above prototypical probability $\rho_{i,k}^t$ with the teacher's output probability $p_{i,k}^t$, to produce a hybrid soft pseudo-label $\eta_{i,k}^t$:

$$\eta_{i,k}^t = \gamma * \rho_{i,k}^t + (1-\gamma) * p_{i,k}^t \tag{10}$$

where γ is a fuse factor.

Since the trained teacher has the general semantic knowledge of classes, the $p_{i,k}^t$ can be regarded as a priori knowledge, to improve the quality of pseudo labeling, which shows appealing advantages in previous works (Li, Xiong, and Hoi 2021; Zhang et al. 2021).

Note that, the teacher network's output $p_{i,k}^t$ remains fixed as training proceeds. The reason we choose $p_{i,k}^t$ instead of the updating probability $q_{i,k}^t$ of the student, is to avoid the degenerate solution, resulting from the simultaneous update of features and labels throughout the self-training. Subsequently, we use the hybrid η_i^t instead of ρ_i^t to produce pseudo hard label. To this end, the student can be trained by the traditional self-training loss (Zou et al. 2018):

$$\mathcal{L}_{CE}(\Theta_{stu}) = -\sum_{x \in D_t^x} \sum_{i=1}^{L} \xi(\eta_i^t) log(q_i^t) \tag{11}$$

where q_i^t denotes the probability distribution produced via the classifier of the student network for x_i^t.

Based on the above, the student network can benefit from two aspects (Fig 2) : knowledge distillation and self-training. A very straightforward issue is that the student network may not be competent at an early stage to undertake effective self-training . To guarantee that the student network can learn the shared class semantic for self-training at the early stage , we follow a cumulative learning strategy (Zhou et al. 2020) to gradually shift the model's learning focus from knowledge distillation to self-training using the control parameter α :

$$\alpha = 1 - (\frac{e}{E_{max}})^2 \tag{12}$$

where E_{max} is the number of total training epochs, and e is the current epoch. The α automatically decreases from 1 to 0 with increasing epoch.

Finally, the loss $\mathcal{L}(\Theta_{stu})$ for the student network can be expressed as:

$$\mathcal{L}(\Theta_{stu}) = (1-\alpha)\mathcal{L}_{CE}(\Theta_{stu}) + \alpha \mathcal{L}_{KD}(\Theta_{stu}) \tag{13}$$

Experiments and Analysis

Datasets

We adopt three widely-used benchmark datasets for experiments: **CoNLL-2002** (Spanish and Dutch) (Sang 2002), **CoNLL-2003** (English and German) (Sang and Meulder 2003), and **Wikiann** (English, Arabic, Hindi and Chinese) (Pan et al. 2017). Each language-specific dataset has the standard training, development, and evaluation sets.The statistics for all datasets are shown in Table 1.

Following previous works (Wu et al. 2020a,c), we apply word-piece (Wu et al. 2016) to tokenize the sentences into sub-words, which then be marked by the BIO scheme. The data are annotated with four different entity types: PER (Persons), LOC (Locations), ORG (Organizations), and MISC (Miscellaneous). For all experiments, English is regarded as the source language and others as the target language respectively. Note that, CoNLL-2002/2003 share a common English dataset as source data. Moreover, we train the model on the source language training set, validate the model on the source language development set, and evaluate the learned model on the target language test set to simulate the zero-resource cross-language NER scenario.

Implementation Details

We adopt the pre-trained mBERT(Pires et al. 2019) as the feature extractor. Following previous works (Wu et al. 2020a,c), we use the token-level F1 score as the evaluation metric. For all experiments, we use the Adam optimizer (Kingma and Ba 2015) with learning rate = 5e-5 for teacher network and 1e-5 for student network, batch size = 128, maximum sequence length = 128, and the dropout = 0.5 empirically. We utilize the grid search technology to obtain the optimal super-parameters, including the moving average coefficient λ selected from $\{0.001, 0.005, 0.0001, 0.0005\}$, the contrastive learning temperature τ_1 selected from 0.5 to 0.9, the softmax temperature τ_2 selected from 0.5 to 0.9, and the fuse factor γ selected from 0.7 to 0.9.

Following the previous work (Wu and Dredze 2020), we only consider the first sub-word tokenized by word-piece in our loss function and freeze the parameters of the embedding

Datasets	Language	Type	Train	Dev	Test
Conll2003	English (en)	Sentence	14,987	3,466	3,684
		Entity	23,499	5,942	5,648
	German (de)	Sentence	12,705	3,068	3,160
		Entity	11,851	4,833	3,673
Conll2002	Spanish (es)	Sentence	8,323	1,915	1,517
		Entity	18,798	4,351	3,558
	Dutch (nl)	Sentence	15,806	2,895	5,195
		Entity	13,344	2,616	3,941
Wikiann	English (en)	Sentence	20,000	10,000	10,000
		Entity	27,931	14,146	13,958
	Arabic (ar)	Sentence	20,000	10,000	10,000
		Entity	22,500	11,266	11,259
	Hindi (hi)	Sentence	5,000	1,000	1,000
		Entity	6,124	1,226	1,228
	Chinese (zh)	Sentence	20,000	10,000	10,000
		Entity	25,031	12,493	12,532

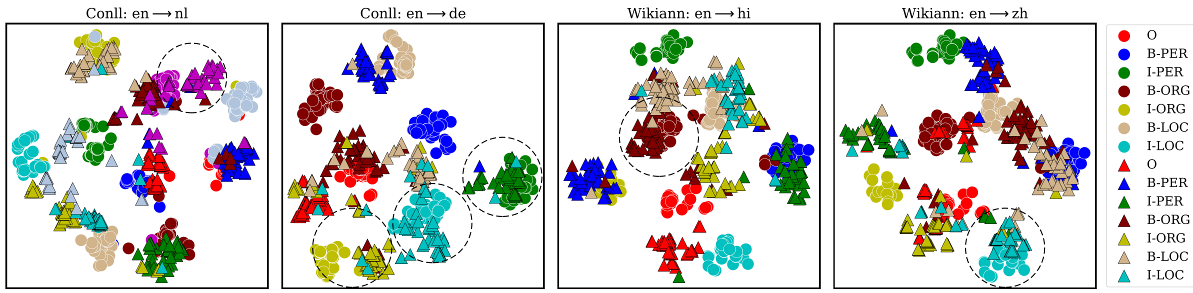
Table 1: Statistics of the datasets.

Method	es	nl	de	Avg
Ni and Dinu (2017)	65.10	65.40	58.50	63.00
Mayhew et al. (2017)	65.95	66.50	59.11	63.85
Xie et al. (2018a)	72.37	71.25	57.76	67.13
Wu and Dredze (2019)	74.50	79.50	71.10	75.03
Moon et al. (2019)	75.67	80.38	71.42	75.82
Wu et al. (2020c)	76.75	80.44	73.16	76.78
Wu et al. (2020a)	76.94	80.99	73.22	77.02
UniTrans (Wu et al. 2020b)	77.30	81.20	73.61	77.37
RIKD(Liang et al. 2021)	77.84	82.46	75.48	78.59
AdvPicker(Chen et al. 2021)	79.00	**82.90**	75.01	78.97
ProKD (Ours)	**79.53**	82.62	**78.90**	**80.35**

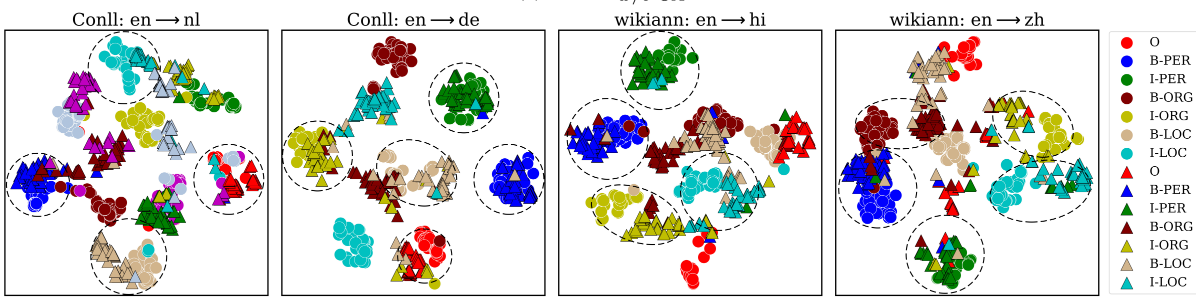
Table 2: Result comparison on Conll2002&2003.

Method	ar	hi	zh	Avg
Wu and Dredze (2020)	42.30	67.60	-	-
Wu et al. (2020a)	43.12	69.54	48.12	53.59
RIKD (Liang et al. 2021)	45.96	70.28	50.40	55.55
ProKD (Ours)	**50.91**	**70.72**	**51.80**	**57.81**

Table 3: Result comparison on Wikiann.



(a) ProKD$_{w/o\ \mathrm{CA}}$



(b) ProKD

Figure 3: Our ProKD shows superiority over ProKD$_{w/o\ \mathrm{CA}}$ with more classes aligned correctly. The circles (●) and triangles (▲) indicate sample representations of the source and target languages, respectively. Dashed circles indicate that samples from different languages belonging to the same class are correctly aligned.

Conll 2022 & 2003				
Method	es	nl	de	average
ProKD	**79.53**	**82.62**	**78.90**	**80.35**
ProKD$_{w/o\ \mathrm{CA}}$	77.46	80.34	77.36	78.25 (-2.10)
ProKD$_{w/o\ \mathrm{ST}}$	77.85	80.69	77.85	78.80 (-1.55)
ProKD$_{w/o\ \mathrm{PK}}$	79.35	82.00	78.56	79.97 (-0.38)
ProKD$_{w/o\ \mathrm{CL}}$	79.42	82.20	78.63	80.08 (-0.27)
Wikiann				
Method	ar	hi	zh	average
ProKD	**50.91**	**70.72**	**51.80**	**57.81**
ProKD$_{w/o\ \mathrm{CA}}$	48.88	69.02	49.57	55.82 (-1.99)
ProKD$_{w/o\ \mathrm{ST}}$	49.61	69.57	50.12	56.43 (-1.38)
ProKD$_{w/o\ \mathrm{PK}}$	50.45	70.41	51.33	57.40 (-0.41)
ProKD$_{w/o\ \mathrm{CL}}$	50.73	70.66	51.56	57.65(-0.16)

Table 4: Ablation study on different factors.

layer and the bottom three layers of the mBERT model. Additionally, our approach is implemented using PyTorch, and all calculations are done on NVIDIA Tesla V100 GPU.

Performance Comparison

We compare the proposed approach with several previous approaches, including three translation-based approaches: **Mayhew et al. (2017)**, **Ni and Dinu (2017)**and **Xie et al. (2018a)** , three direct tranfer-based approaches: **Wu and Dredze (2019)**, **Moon et al. (2019)** and **Wu et al. (2020c)**, and four knowledge distillation-based approaches: **Wu et al. (2020a)**, **UniTrans (Wu et al. 2020b)**, **RIKD(Liang et al. 2021)** and **AdvPicker(Chen et al. 2021)**.

The results are presented in Table 2 and Table 3 ,where the baseline and SOTA experimental results are from their original papers. As observed, our method achieves the best results on most the datasets. For Conll2002/2003, compared with the two competitive knowledge distillation-based methods, RIKD and AdvPicer, our approach improves on the average F1 by 1.76% and 1.38%, respectively. For Wikiann, our method outperforms the RIKD by 2.26% on average. Especially, for German(de) language, we obtain an F1 value of 78.9%, which is 3.42% higher than the best result of the RIKD. And for Arabic(ar) language, our method achieves the best F1 value of 50.91%, with an improvement of 4.95% than RIKD. Analytically, RIKD and AdvPicer leverage adversarial learning and reinforcement learning to select target data for distillation, respectively, and the selected data tends to be consistent with the source language in feature distribution. Consequently, the student network learning on this data fail to effectively acquire the target language knowledge, resulting in insufficient generalization on the target language. Contrastly, our model uses the prototypical self-training to enhance the student network's ability to learn the target language, thus performing well on the target language.

Ablation Study

To investigate the contributions of different factors, we conduct ablation experiments with four variant models: (1) **ProKD**$_{w/o\ \mathrm{CA}}$ removes the prototypical class-wise alignment from the teacher network. (2) **ProKD**$_{w/o\ \mathrm{ST}}$ wipes out

	test text	source	target
#1	ProKD$_{w/o\ \mathrm{ST}}$: Para UGT de Madrid [I-ORG], la decisión ...	66.67% [I-ORG]	32.16% [I-ORG]
Spanish	ProKD: Para UGT de Madrid [B-LOC], la decisión ...	16.67% [B-LOC]	59.73% [B-LOC]
#2	ProKD$_{w/o\ \mathrm{ST}}$: Sozialdezernent Martin [I-PER] Berg war ...	53.19% [B-PER]	84.38% [B-PER]
German	ProKD : Sozialdezernent Martin [B-PER] Berg war ...	46.80% [I-PER]	9.38% [I-PER]
#3	ProKD$_{w/o\ \mathrm{ST}}$: we dit najaar ... met Vandenbroucke[I-PER] ...	100.00% [I-PER]	33.33% [I-PER]
Dutchc	ProKD:we dit najaar ... met Vandenbroucke [B-PER]...	0.00% [B-PER]	66.67% [B-PER]

Table 5: Case Study. The ProKD can learn language-specific knowledge with self-training, which helps the model to rectify incorrect predictions to correct ones.

the prototypical self-training from the student network. (3) **ProKD**$_{w/o\ \mathrm{PK}}$ does not use the prior knowledge from the teacher network in self-training process. (4) **ProKD**$_{w/o\ \mathrm{CL}}$ cuts out the cumulative learning scheme and adopts the parameter $\alpha = 0.5$ in loss function (Equation 13) for student network. As shown in Table 4, the average F1 value of ProKD$_{w/o\ \mathrm{CA}}$ decrease by 2.1% compared to ProKD on Conll 2002 & 2003. This indicates that class-level alignment effectively improves the model's generalization, as class alignment forces the teacher network to learn language-independent knowledge from source and target languages. The performance of the ProKD$_{w/o\ \mathrm{ST}}$ in F1 score drops by 1.55% compared to ProKD, which well validates the effectiveness of the self-training to acquire the target language-specific knowledge. For ProKD$_{w/o\ \mathrm{PK}}$, the slight drop in F1 results compared to the ProKD suggests that incorporating prior knowledge of the teacher network can enhance the quality of pseudo labels during self-training. Also, ProKD$_{w/o\ \mathrm{CL}}$ yields a slight drop in F1 values, which proves that the knowledge distillation learning should be performed first and then the self-training. The above experimental phenomena can also be observed on the Wikiann dataset.

Visualizing the Token Sample Representations

To demonstrate that our ProKD can achieve class-level feature alignment, we randomly select 50 token samples for each class from the source and target languages and feed them to the teacher networks of ProKD and ProKD$_{w/o\ \mathrm{CA}}$ to obtain token-level representations, respectively. Note that, the teacher network of ProKD$_{w/o\ \mathrm{CA}}$ degenerates to a vanilla mBERT when removing the prototypical class-wise alignment. We then visualize these representations using the T-SNE (Van der Maaten and Hinton 2008) and show the results for the four target languages in Figure 3. As shown, the feature representations of source and target languages from ProKD$_{w/o\ \mathrm{CA}}$ are distributed differently and inconsistent due to languages gap. Many target language examples of one class are incorrectly aligned to the source language examples of a different class, thus causing confusion and hindering the model's performance. By contrast, our approach ProKD shows superiority over ProKD$_{w/o\ \mathrm{CA}}$ with more classes aligned correctly. For example, when performing a cross-lingual NER from English(en) to Chinese(zh), the ProKD$_{w/o\ \mathrm{CA}}$ aligns source and target features for just one class, ILOC, while our model achieves feature alignment on five classes. We argue that a model aligning features across multiple classes can capture more shared class features across languages, which is essential for generalizing the model to unknown target languages.

Case Study

In this part, we present a case study to show that our model can learn target language-specific knowledge through self-training. We compare the prediction results of the ProKD$_{w/o\ \mathrm{ST}}$ with our ProKD for the target language test data, as shown in Table 5. In example 1, the ProKD$_{w/o\ \mathrm{ST}}$ model incorrectly predicts "Madrid" as "I-ORG" because 66.67% of the "Madrid" tokens in English dataset are annotated as "I-ORG". The teacher network trained with this English data would distill this label preference to the student network, resulting in the student network of ProKD$_{w/o\ \mathrm{ST}}$ tending to make incorrect predictions. In contrast, our model captures the label preferences of the target language by a prototypical self-training learning mechanism. In the same example 1, 59.73% of "Madrid" tokens in target Spanish language are labeled as "I-ORG". Our model can produce accurate predictions due to its intimate familiarity with the target language-specific knowledge. We observe the same phenomenon in examples 2 and 3.

Conclusion

This paper presents a knowledge distillation-based network ProKD for zero-resource cross-lingual NER. The ProKD proposes a contrastive learning-based prototype alignment approach to boost the teacher network's capacity to capture language-independent knowledge. In addition, the ProKD introduces the prototypical self-training method to improve the student network's capacity to grasp language-specific of target knowledge. The experiments on six target languages illustrate the effectiveness of the proposed approach.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work is funded by the National Key Research and Development Program of the Ministry of Science and Technology of China (No.

2021YFB1716201). Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing.

References

Cao, S.; Shi, J.; Yao, Z.; Lv, X.; Yu, J.; Hou, L.; Li, J.; Liu, Z.; and Xiao, J. 2022. Program Transfer for Answering Complex Questions over Knowledge Bases. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 8128–8140. Association for Computational Linguistics.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Chen, W.; Jiang, H.; Wu, Q.; Karlsson, B.; and Guan, Y. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proc. of ACL*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Jain, A.; et al. 2019. Entity Projection via Machine Translation for Cross-Lingual NER. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *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*, 1083–1092. Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Li, J.; Xiong, C.; and Hoi, S. C. H. 2021. MoPro: Webly Supervised Learning with Momentum Prototypes. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Liang, S.; Gong, M.; Pei, J.; Shou, L.; Zuo, W.; Zuo, X.; and Jiang, D. 2021. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. In *Proc. of KDD*.

Liu, C.; Tao, C.; Feng, J.; and Zhao, D. 2022. Multi-Granularity Structural Knowledge Distillation for Language Model Compression. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 1001–1011. Association for Computational Linguistics.

Liu, L.; Ding, B.; Bing, L.; Joty, S. R.; Si, L.; and Miao, C. 2021. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 5834–5846. Association for Computational Linguistics.

Ma, J.; Ballesteros, M.; Doss, S.; Anubhai, R.; Mallya, S.; Al-Onaizan, Y.; and Roth, D. 2022a. Label Semantics for Few Shot Named Entity Recognition. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, 1956–1971. Association for Computational Linguistics.

Ma, T.; Jiang, H.; Wu, Q.; Zhao, T.; and Lin, C. 2022b. Decomposed Meta-Learning for Few-Shot Named Entity Recognition. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, 1584–1596. Association for Computational Linguistics.

Mayhew, S.; et al. 2017. Cheap Translation for Cross-Lingual Named Entity Recognition. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2536–2545. Association for Computational Linguistics.

Moon, T.; Awasthy, P.; Ni, J.; and Florian, R. 2019. Towards Lingua Franca Named Entity Recognition with BERT. *CoRR*.

Ni, J.; and Dinu, G. 2017. Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. In *Proc. of ACL*.

Pan, X.; Zhang, B.; May, J.; Nothman, J.; Knight, K.; and Ji, H. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1946–1958. Association for Computational Linguistics.

Peng, S.; Huang, X.; Lin, Z.; Ji, F.; Chen, H.; and Zhang, Y. 2019. Teacher-Student Framework Enhanced Multi-domain Dialogue Generation. *CoRR*, abs/1908.07137.

Pfeiffer, J.; Vulic, I.; Gurevych, I.; and Ruder, S. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 7654–7673. Association for Computational Linguistics.

Pires, T.; et al. 2019. How Multilingual is Multilingual BERT? In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy,* *July 28- August 2, 2019, Volume 1: Long Papers*, 4996–5001. Association for Computational Linguistics.

Qiu, Z.; Zhang, Y.; Lin, H.; Niu, S.; Liu, Y.; Du, Q.; and Tan, M. 2021. Source-free Domain Adaptation via Avatar Prototype Generation and Adaptation. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 2921–2927. ijcai.org.

Sang, E. F. T. K. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In Roth, D.; and van den Bosch, A., eds., *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*. ACL.

Sang, E. F. T. K.; and Meulder, F. D. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Daelemans, W.; and Osborne, M., eds., *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, 142–147. ACL.

Snell, J.; et al. 2017. Prototypical Networks for Few-shot Learning. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4077–4087.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van Nguyen, M.; Nguyen, T. N.; Min, B.; and Nguyen, T. H. 2021. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5414–5426.

Weng, R.; Yu, H.; Huang, S.; Cheng, S.; and Luo, W. 2020. Acquiring Knowledge from Pre-Trained Model to Neural Machine Translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 9266–9273. AAAI Press.

Wu, Q.; Lin, Z.; Karlsson, B.; Lou, J.; and Huang, B. 2020a. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In *Proc. of ACL*.

Wu, Q.; Lin, Z.; Karlsson, B. F.; Huang, B.; and Lou, J. 2020b. UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data. In *Proc. of IJCAI*.

Wu, Q.; Lin, Z.; Wang, G.; Chen, H.; Karlsson, B. F.; Huang, B.; and Lin, C. 2020c. Enhanced Meta-Learning for Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proc. of AAAI*.

Wu, S.; and Dredze, M. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *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*, 833–844. Association for Computational Linguistics.

Wu, S.; and Dredze, M. 2020. Do Explicit Alignments Robustly Improve Multilingual Encoders? In *Proc. of EMNLP*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Xie, J.; Yang, Z.; Neubig, G.; Smith, N. A.; and Carbonell, J. G. 2018a. Neural Cross-lingual Named Entity Recognition with Minimal Resources. In *Proc. of EMNLP*.

Xie, S.; Zheng, Z.; Chen, L.; and Chen, C. 2018b. Learning Semantic Representations for Unsupervised Domain Adaptation. In *Proc. of ICML*.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022. VGSE: Visually-Grounded Semantic Embeddings for Zero-Shot Learning. *CoRR*, abs/2203.10444.

Yang, Z.; Shou, L.; Gong, M.; Lin, W.; and Jiang, D. 2020. Model Compression with Two-stage Multi-teacher Knowledge Distillation for Web Question Answering System. In *Proc. of WSDM*.

Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; and Wen, F. 2021. Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 12414–12424. Computer Vision Foundation / IEEE.

Zhou, B.; Cui, Q.; Wei, X.; and Chen, Z. 2020. BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9716–9725. Computer Vision Foundation / IEEE.

Zou, Y.; Yu, Z.; Kumar, B. V. K. V.; and Wang, J. 2018. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-training. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, 297–313. Springer.