Scilight

Article

Improving Named Entity Recognition Using Multi-Domain Knowledge Transfer

Ziyan Wang [1], Peng Zhu [1,*], Yuanzhe Chen [2], and Qinyuan Liu [1]

[1] School of Computer Science and Technology, Tongji University, Shanghai 201804, China
[2] School of Data Science and Engineering, East China Normal University, Shanghai 200062, China
* Correspondence: pengzhu@tongji.edu.cn

Received: 21 April 2026; Revised: 12 July 2026; Accepted: 6 November 2026; Published: 13 November 2026

Abstract: Neural networks have been widely used for named entity recognition (NER) and have shown state-of-the-art performance. However, in real applications, we often need to consider new types of entities in new domains where we do not have sufficient annotated data, which inevitably leads to a degradation in NER model performance. To this end, we propose an end-to-end domain adversarial named entity recognition (DANER) model. Unlike previous rule-based methods, DANER explicitly addresses NER-specific token-level context shifts by deeply integrating a gradient reversal layer with a sequence tagging architecture. Extensive experiments on our newly constructed FIND-2026, to our knowledge, the first large-scale Chinese NER corpus explicitly enriched with fine-grained domain annotations, demonstrate the superiority of our proposed model over six state-of-the-art methods. Moreover, on two public datasets (MSRA and Weibo), our approach achieves remarkable zero-shot knowledge transfer capabilities, yielding an absolute F-score gain of more than 22 percentage points compared with non-domain baselines without utilizing any target-domain labels.

Keywords: named entity recognition; domain adversarial; multi-domain embeddings; domain-invariant knowledge

1. Introduction

Named Entity Recognition (NER) [1] is defined as the extraction of a contiguous sequence of textual tokens, which represents the name of an object of a specified class, such as a company, location, or person. It plays a vital role in the overall task of Information Extraction (IE) and serves as an intermediate step for subsequent IE tasks, like entity linking and relation extraction [2]. Current state-of-the-art methods for NER usually use deep learning-based algorithms [3,4], e.g., feed-forward neural networks (FNNs) or recurrent neural networks (RNNs) [5], and build name taggers from annotated data with accompanying entity labels. Such models generalize well to new entities based on the features automatically learned from the context. However, a neural-based NER system could still get lower performance if its hidden feature representations cannot be learned adequately, which happens frequently when there is not enough annotated data in a new domain.

In real-world applications, NER systems frequently encounter out-of-domain text containing new types of entities lacking annotated data. For instance, a model trained on a labeled corpus from the industrial and financial sectors often experiences a significant performance drop when applied to the Information Technology (IT) domain. Acquiring fine-grained entity annotations for every new domain is highly expensive and non-scalable. Fortunately, obtaining coarse-grained domain categories (e.g., classifying a document as “Finance” or “IT”) is much more feasible, as this process can be easily accelerated by semi-supervised methods or domain experts.

However, even with available domain categories, transferring knowledge across domains remains non-trivial due to inherent feature distribution shifts. As illustrated in Figure 1, standard word embedding algorithms, such as word2vec [6], GloVe [7], and BERT [8], tend to represent entities of different domains separately, as the same domain entities are more likely to co-occur in corpus context and are subsequently embedded closer. These may cause a shift among domain distributions. Thus, to get a good recognition accuracy for new domains and also for the conditions of new entity types, we need to build mappings between the source and the target domains, so that the model learned from the source can also be applied to the target when composed with the learned mapping between domains.

Recently developed domain adaptation methods [9] attempt to mitigate the harmful effects of domain shift. They learn deep transformations that map both domains into a common feature space [10], which have shown good progress in image [11], video [12] and bioinformation [13] tasks. Conventional domain adaptation algorithms [14,15] on named entity recognition are mainly powered by rule-based bootstrap approaches. These rules are hard to maintain, and they would inevitably involve semantic drift [16,17] during bootstrapping.

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher’s Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

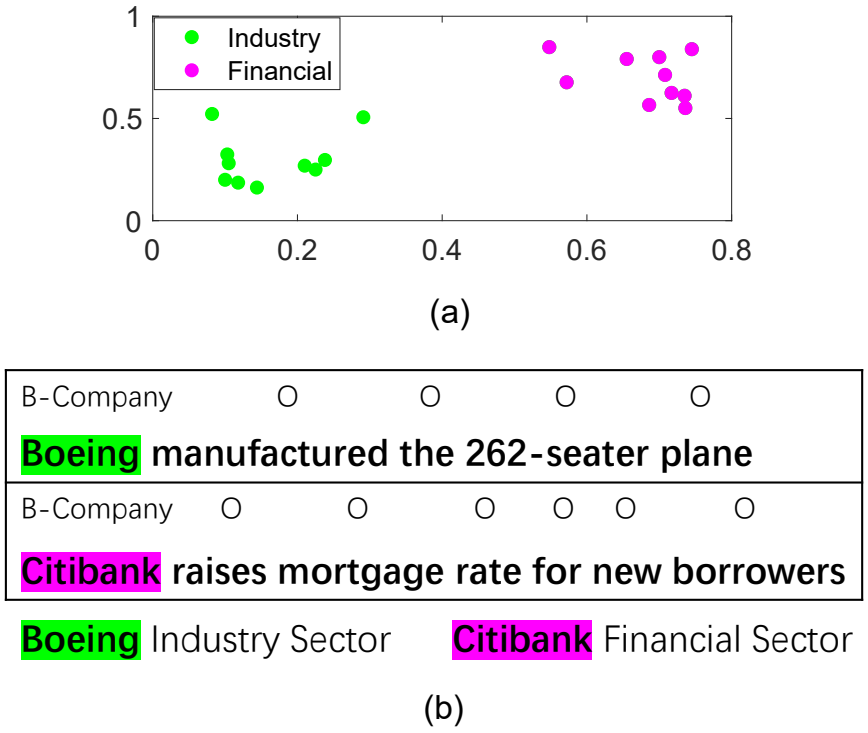


Figure 1. (**a**) We visualize the word embeddings of typical companies from industry and financial sectors. These co-occurrence (or context) based representations are the main causes of feature shifts among different domains. (**b**) An example of NER annotation with domain categories, which could easily be labeled, clustered, or even directly linked to S&P 500 Sectors. Our goal is to embed domain adaptation into the process of learning representation, so that the named entity recognitions are made based on the features that are both discriminative and invariant to the change of domains, i.e., have very similar distributions in the source and the target domains.

Therefore, in this paper, we propose a domain adversarial recurrent neural network for named entity recognition (DANER). Instead of aligning global features like in computer vision, we uniquely tailor the adversarial mechanism to address the token-level feature distribution shifts specific to sequence labeling. We focus on learning features that combine discriminability and domain invariance, which is achieved by jointly optimizing the underlying features as well as two discriminative classifiers operating on these features: (1) the named entity predictor, a conditional random fields (CRF) [18] layer, that predicts entity labels and is used during training and test time; (2) the domain classifier that discriminates among multiple source domains during training. The parameters of the underlying deep feature mapping are optimized to minimize the loss of the CRF classifier and to maximize the loss of the domain classifier. Since existing publicly available datasets for named entity recognition lack domain categories, we constructed a massive manually annotated Financial Domain (FIND-2026) dataset specifically for our experimental evaluation. Extensive experiments demonstrate the effectiveness of our proposed methods. In summary, the main contributions of this paper are:

- We propose DANER, an end-to-end domain adversarial neural network tailored for cross-domain named entity recognition. To overcome the semantic drift of traditional rule-based methods and address NER-specific token-level context shifts, DANER deeply integrates a gradient reversal layer (GRL) with a sequence tagging architecture, jointly optimizing for entity discriminability and domain invariance.
- We construct FIND-2026, a manually annotated financial NER dataset comprising over 14,000 entities across 9 industry sectors. As the first large-scale Chinese NER corpus with explicit domain annotations, it breaks the data barrier in current cross-domain NER research and serves as a foundational testbed for adversarial evaluation.
- Extensive experiments demonstrate the superiority of DANER on the FIND-2026 dataset. It achieves substantial absolute improvements over traditional baselines and provides consistent, orthogonal gains over strong pre-trained models (e.g., BERT, GPT-NER). Furthermore, it achieves remarkable zero-shot knowledge transfer, yielding absolute F-score gains of over 22 percentage points on public datasets (MSRA and Weibo) without utilizing any target-domain entity labels.

2. Related Works

We summarize the related works in two main areas: named entity recognition and transfer learning on domain adaptation.

Named entity recognition Named entity recognition is typically regarded as a kind of sequence labeling problem in the literature. Therefore, standard feature-based classification approaches such as conditional random fields (CRFs), hidden Markov models (HMMs) and maximum entropy classifiers can be naturally employed to build a name tagger. Despite the effectiveness of feature engineering, it is labor-intensive and unable to discover the discriminative or explanatory factors of data [19]. To handle this problem, some recent studies [3,20] used neural network methods and got promising results. The representative approaches include BiLSTM-CRF [21], CNN-CRF [22] and LSTM-CNN-CRF [23]. Furthermore, character-based representations [8,24] have been proven to be effective in capturing orthographic and morphological evidence. Ref. [9] proposed a transfer learning method that adapts a high-resource English clinical NER model to low-resource languages and domains using only small amounts of in-domain annotated data. Ref. [25] designed an efficient named entity recognition framework for marine domain texts to improve the recognition of long and nested entities, and the attention mechanism enhanced the model's ability to recognize complex entity structures in coral reef ecosystem texts. Our architecture is based on the success of improving representations and further adding domain adversarial networks to capture features that are both discriminative and invariant to the change of domains.

Transfer learning on domain adaptation Domain adaptation [26] is transfer learning under the shift between training and test distributions. A rich line of approaches to domain adaptation aims to bridge the gap between the source and target domains by learning domain-invariant feature representations. Such deep representations allow the classifier learned from the source domain to be applied to the target domain effectively. Most works [27,28] focused on minimizing the maximum mean discrepancy (MMD) metric in the new feature space by matrix factorization or adding regularization into the neural networks. Ref. [29] argued that the source classifier and target classifier cannot be shared even if the distribution gap is greatly reduced with the new feature representations. Hence, they introduced the residual function to learn the difference between the source and target classifiers explicitly. Inspired by the idea of adversarial learning [10], some of the latest works [30–33] proposed a discriminator network to supervise the learning of the domain-invariant features. However, all these methods are typically applied to computer vision and data mining applications [34–36]. Ref. [37] developed an effective method called Knowledge Transfer Learning via Dual Density Sampling (KTL-DDS), which consists of Dual Density Sampling (DDS), Consistent Maximum Mean Discrepancy (CMMD) and Knowledge Dissemination (KD). Ref. [38] proposed a novel multi-source transfer learning energy prediction model based on long short-term memory (LSTM) and multi-kernel maximum mean difference (MK-MMD) domain adaptation, which was used for the short-term energy prediction of different types of buildings lacking historical data. To the best of our knowledge, we are the first to apply the domain adaptation method for transferring sequence embeddings from one sector to another for named entity recognition. We also exploit multi-source data for discriminative feature learning.

3. Methodology

In this section, we first describe the background of sequence models, which are the backbone of our approach. Afterward, we present the domain network to learn both discriminative and transferable feature representations for domain adaptation. Lastly, we introduce the joint optimization strategy that leaves the input unchanged during forward propagation and reverses the gradient during backpropagation.

3.1. Basic Model: LSTM-CRF

The LSTM-CRF model was originally introduced by [5], which takes a sequence of elements as the input and outputs a sequence of category labels corresponding to the input sequence. The idea has been successfully applied in POS Tagging, Chunking, and NER [39].

We build high-level neural features from the input character sequence by a bi-directional LSTM [21]. The resulting features are combined and then fed into an output CRF layer for decoding. The architecture of this backbone network is shown in the upper part of Figure 2. The feature extractor takes a sequence of vectors $X = \{x_1, \cdots, x_i, \cdots, x_n\}$ as input and returns a hidden sequence $H = \{h_1, \cdots, h_i, \cdots, h_n\}$ by using a bi-directional LSTM (Bi-LSTM) structure together with a linear transformation over the inputs. CRF tagging module is applied over H to obtain the final output named entity labels $y = \{y_1, \cdots, y_i, \cdots, y_n\}$. It defines the score function in Equation (1):

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} H_{i, y_i} \tag{1}$$

where A is a matrix of transition scores such that $A_{i,j}$ represents the score of a transition from the tag i to tag j. y_0 and y_n are the start and end tags of a sentence. Matrix A is therefore a square matrix of size $k+2$, and k is the number of tags. A softmax over all possible tag sequences yields the probability of sequence y in Equation (2):

$$p(y|X) = \frac{\exp(s(X,y))}{\sum_{\tilde{y}\in Y_x} \exp(s(X,\tilde{y}))} \tag{2}$$

where Y_x represents all possible tag sequences for the input sequence X.

The parameter θ_y in the CRF layer G_y is optimized by maximizing the log-probability of the correct tag sequence in Equation (3):

$$\mathcal{L}_y = s(X,y) - \log\left(\sum_{\tilde{y}\in Y_x} \exp(s(X,\tilde{y}))\right) \tag{3}$$

Furthermore, some work [24] incorporated character-level structure into word representation. Each input vector x_i consists of two parts, a pre-trained word-level representation w_i and a task-related character-level representation c_i. They adopted a bidirectional LSTM to capture information in both forward and backward directions and concatenated the outputs of these two LSTMs as c_i, which could also be integrated into our proposed method.

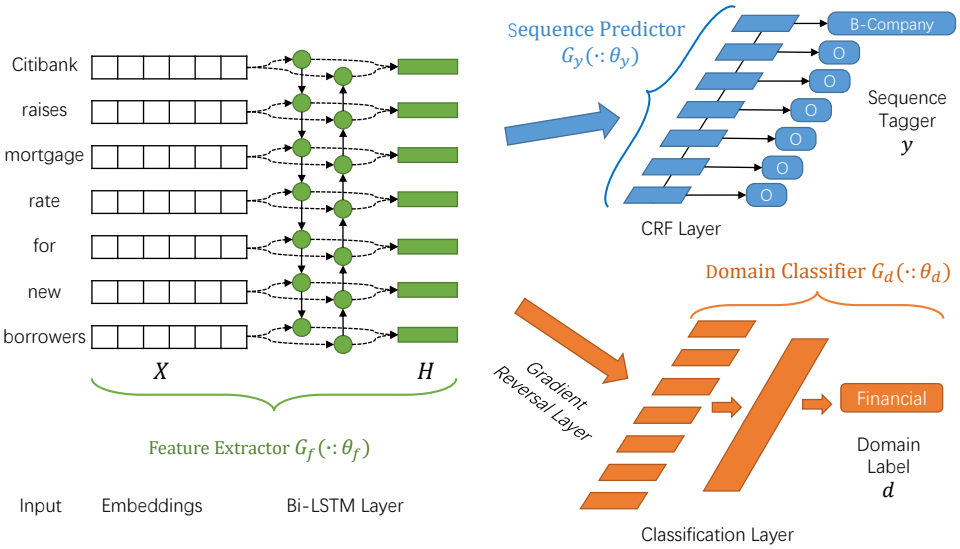


Figure 2. The proposed domain adversarial named entity recognition architecture includes a Bi-LSTM layer (green), a CRF-based sequence predictor layer (blue), and a shallow neural network-based domain classifier (orange). Inputs (word embeddings) are first fed to the Bi-LSTM layer for feature extraction, then simultaneously linked to a sequence predictor for named entity recognition and a domain network for named entity domain classification. Domain adaptation is achieved by adding a domain classifier connected to the feature extractor (Bi-LSTM) by a gradient reversal layer that multiplies the gradient by a certain negative constant during the back-propagation-based training. Then, the training proceeds standardly and minimizes the sequence prediction loss and the domain classification loss. Gradient reversal ensures that the feature distribution over different domains is made as indistinguishable as possible for the domain classifier, thus resulting in domain-invariant features.

3.2. Domain Network

Domain network guides the model to learn transferable feature representations H for domain adaptation. Applying Bi-LSTM and CRF trained on embeddings from the source domain may suffer from poor performance in the target domain due to the feature distribution shift between domains. Many works [40,41] realize this new space by minimizing maximum mean discrepancy (MMD) [27], a metric for dissimilarity between two feature distributions. However, this process is typically non-trivial because the source and target feature distributions are often high-dimensional and constantly changing during the training phase.

Thus, to explicitly learn transferable feature representations H, we introduce a domain discriminator network G_d. G_d acts as an adversary against the feature extractor (Bi-LSTM), supervising it towards learning features that are discriminative for NER but invariant across domains. Specifically, G_d takes the hidden representations H produced by the Bi-LSTM and attempts to correctly classify which domain the input text belongs to. To achieve

this, we design G_d as a multi-layer perceptron (MLP) consisting of two fully connected feed-forward layers and a softmax output layer. Given an output $H = G_f(X, \theta_f)$ from the Bi-LSTM, the forward pass of the domain network is formally defined in Equation (4):

$$\begin{aligned} Z_{d1} &= \sigma(W_{d1}H + b_{d1}) \\ Z_{d2} &= \sigma(W_{d2}Z_{d1} + b_{d2}) \\ \widehat{d} &= \frac{\exp(\tilde{z}_{d2})}{\sum_z \exp(z_{d2})} \end{aligned} \tag{4}$$

where the W and b terms are weight matrices and bias vectors respectively, w_{d3} denotes the neuron weights of the output layer and $\widehat{d}$ is the predicted domain label.

The parameters θ_d in G_d are optimized by maximizing the cross-entropy loss $\mathcal{L}_d$ for domain classification in Equation (5):

$$\mathcal{L}_d = -\sum_{i=1}^{T} d_i \log \widehat{d}_i \tag{5}$$

where d is the ground-truth domain label indicating the named entity is from the category.

3.3. Optimization

The proposed DANER model involves three sets of parameters for its submodules, namely θ_f from the Bi-LSTM feature extractor, θ_y from the CRF sequence predictor, and θ_d from the domain classifier. These parameters are optimized jointly by minimizing $\mathcal{L}_y$ to learn discriminative features for named entity prediction, and maximizing $\mathcal{L}_d$ to learn domain-invariant features for domain adaptation. Hence, the overall loss function $\mathcal{L}$ is defined in Equation (6):

$$\mathcal{L} = \mathcal{L}_y(X, y; \theta_f, \theta_y) - \lambda \mathcal{L}_d(X, d; \theta_f, \theta_d) \tag{6}$$

where the hyperparameter λ controls the trade-off between sequence prediction and domain classification.

It is worth noticing that the loss for the domain network cannot be optimized by the stochastic gradient descent (SGD) method directly. This is because the parameters θ_d in G_d need to be updated to enhance the classification ability. When the gradients $\frac{\partial \mathcal{L}_d}{\partial \theta_f}$ are back-propagated from G_d to G_f, they are reversed and the reversed gradients drive Bi-LSTM to learn new feature representations that are even harder to be distinguished by G_d, i.e., the parameters θ_f are updated to increase domain classification error. To solve the problem, we adopt a gradient reversal layer (GRL) [42] to optimize this network. During forward propagation, GRL acts as an identity mapping, while it multiplies each gradient from the subsequent level by $-\lambda$ and passes the result to the preceding layer during backward propagation. As shown in Figure 2, we insert GRL between Bi-LSTM and the domain network, based on which we can optimize the domain adversarial NER model using the standard SGD process.

In our implementation, we apply the Adam optimizer [43] to update the parameters θ_f, θ_y, θ_d iteratively. The initial learning rate is set to 0.001, λ to 1, the number of hidden units of the LSTM is set to 256, and the batch size is set to 128 by default.

4. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness and efficiency of our proposed methods. We first provide an overview of the datasets and methods that we use in our experiments. We then thoroughly evaluate the model in named entity prediction tasks compared with other baselines. After that, the effects of the domain adversarial layer are experimented with and visualized, respectively. Finally, we provide the case studies of the proposed methods.

4.1. Experimental Settings

4.1.1. Datasets

We collected the FIND-2026 dataset from an industry partner, which crawled news articles related to the CSI-300 index constituent companies over a two-year period (from 1 July 2016 to 30 June 2018). We then divide these named entities into domains according to the Global Industry Classification Standard (GICS) [44]; they are: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Financials, Information Technology (I.T.), Communication Services (C.S.), Utilities, Health Care, Real Estate. As most consumer sector companies in the CSI-300 overlap their business in consumer discretionary and consumer staples, we combine the two into the Consumer sector. We also merge energy into the industrial sector. Thus, there are 9 domains finally with a total of

9769 sentences and 14,230 entities in the FIND-2026 dataset, which are detailed in Table 1.

Table 1. Statistics of training and testing sets in the FIND-2026 dataset.

Sectors	Training and Validation Sets		Test Sets	
	#Entities	#Sentences	#Entities	#Sentences
Materials	1835	1436	-	-
Industrials	2576	2082	-	-
Financials	1096	889	-	-
I.T.	2586	2056	-	-
C.S.	399	324	-	-
Real Estate	562	448	-	-
Utilities	-	-	1479	1174
Consumer	-	-	2217	1696
Health Care	-	-	1480	1194
Total	9054	7235	5176	2534

To ensure the high quality of the FIND-2026 dataset, the crawled news articles related to the CSI-300 index constituent companies underwent a rigorous manual annotation process. Following standard NER practices, our annotation guidelines specifically targeted entities crucial to financial information extraction, including companies (e.g., annotated as "B-Company"), persons, and locations. The annotation was conducted by a team of three graduate students with a strong background in finance and natural language processing, following a two-pass strategy of independent labeling and subsequent cross-validation. Any conflicting annotations during this phase were discussed and resolved by a senior researcher to establish the final ground truth. To quantitatively evaluate the annotation reliability, we measured the Inter-Annotator Agreement, which indicates substantial and highly reliable agreement among the annotators. Finally, to facilitate future research in cross-domain NER and domain adaptation, the FIND-2026 dataset, along with our detailed annotation guidelines, will be made publicly available upon acceptance of this paper.

We randomly choose sentences from 6 of the 9 sectors as training sets, and sentences in the remaining 3 sectors (Utilities, Consumer, Health Care) are utilized as test sets. In our experiments, we aim to validate the effectiveness of integrating domain adversarial training into recurrent neural networks for NER tasks and to evaluate the sequence prediction accuracy of the learned domain-invariant features against other state-of-the-art methods.

To ensure rigorous model evaluation and prevent data leakage, we further randomly divide the 7235 sentences from the 6 source sectors into a training set and a validation set with a ratio of 7.2:1. The validation set is strictly utilized for hyperparameter tuning (e.g., the trade-off parameter λ) and early stopping. Therefore, the overall split ratio of the training, validation, and test sets is approximately 65% (6353 sentences), 9% (882 sentences), and 26% (2534 sentences), respectively. The test set comprises entirely unseen sectors to strictly evaluate the zero-shot cross-domain generalization.

It is crucial to note that our experiments follow a strict multi-source domain generalization (or zero-shot cross-domain) setting. The domain adversarial network is trained exclusively on multiple source domains to learn generalized, domain-invariant representations. No target-domain data (neither labeled nor unlabeled) from the test sectors is exposed to the model during the training or validation phases.

4.1.2. Compared Methods

We compare with the following baseline methods on the FIND datasets.

- LSTM-CRF [21] (denoted as L): a standard bidirectional LSTM and CRF model setup with pre-trained word embeddings (word2vec).
- ID-CNN-CRF [20] (denoted as S) uses iterated dilated convolutional neural networks for named entity recognition with word2vec embeddings.
- Flair [24] (denoted as A): a contextualized embedding method that performs comparably in NER tasks. We utilize Flair embeddings as inputs of two baselines: [24]$^{+L}$ and [24]$^{+S}$.
- BERT [8] (denoted as D^* (indicates methods which utilize external training corpus)): we used the pre-trained representations of BERT and fine-tuned them for the NER task. Then, we reach two strong baselines: [8]$^{*+L}$ and [8]$^{*+S}$.
- GPT-NER [45]: a method aimed at customizing Large Language Models (LLM) for NER. To address the distinction between sequence labeling and text generation tasks, this model guided the LLM to produce a labeled sequence by encapsulating entities with specific tokens.

Our model has several variations, which are:

- DANER^{+L} extends [21] by taking proposed methods.
- DANER^{+S}: an extension of [20] with domain adversarial networks.
- DANER^{+A+L} extends [21] with Flair embeddings and DANER.
- DANER^{+A+S}: an extension of [24]$^{+S}$ with DANER.
- DANER$^{+D^*+L}$ extends [21] with BERT embeddings and domain adversarial networks.
- DANER$^{+D^*+S}$: an extension of [20] with DANER and BERT embeddings.

4.2. Results Analysis

In Table 2, we show the test results of each sector and the total accuracy compared with six state-of-the-art baselines. As we can see, LSTM-CRF [21] is comparable with ID-CNN [20], with improved feature embeddings from word2vec to Flair [24], the accuracy slightly improved. BERT [8] performs much better because we utilize the pre-trained model on an external massive corpus. Rows 8–13 show the results of DANER and some of its variations. Adding DANER consistently improves all baselines on FIND-2026. For standard baselines lacking massive pre-training (e.g., LSTM-CRF, ID-CNN-CRF), the adversarial training yields substantial absolute improvements ranging from 5 to over 15 percentage points. More importantly, when applied to strong pre-trained baselines, DANER$^{+D^*+S}$ pushes the F-Score to 91.42%. While the absolute gain over BERT^{*+S} and GPT-NER is more modest (approx. 0.5 to 1.5 percentage points), this demonstrates that our domain-adversarial network provides orthogonal benefits, effectively capturing domain-invariant features that even strong pre-trained language models may struggle to align across different sectors.

Table 2. Accuracy of named entity recognition in FIND-2026 datasets, where $+L$ means using LSTM-CRF [21]; $+S$ ID-CNN-CRF [20]; $+A$ Flair [24] and $+D$ BERT [8] respectively. We report out the precision (P), recall (R), and F-Score (F), as we can see, DANER improves all baselines from 5% to 15%.

Methods	Utilities			Consumer			Health Care			All		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
LSTM-CRF	87.19	53.67	66.44	79.07	34.91	48.43	79.08	36.81	50.00	81.83	41.34	54.93
ID-CNN-CRF	80.29	59.78	68.53	73.47	38.87	50.84	78.55	53.04	63.32	77.83	49.56	60.56
Flair $^{+L}$	84.38	71.19	77.23	80.39	49.58	61.34	85.49	61.43	71.49	83.51	59.42	69.43
Flair $^{+S}$	77.89	77.58	77.74	73.83	54.21	62.51	81.42	72.94	76.94	77.72	66.63	71.75
BERT $^{*+L}$	91.87	89.13	90.48	92.44	74.61	82.57	93.31	84.97	88.95	92.65	81.95	86.97
BERT $^{*+S}$	90.12	95.51	92.74	90.01	83.93	86.86	90.96	94.04	92.48	90.42	90.29	90.35
GPT-NER	90.65	95.65	93.08	90.12	84.24	87.09	91.05	94.28	92.63	90.85	91.32	91.08
DANER^{+L}	84.01	56.39	67.48	77.68	36.01	49.21	77.69	41.01	53.67	80.30	44.02	56.86
DANER^{+S}	74.39	66.71	70.34	64.98	40.44	49.86	74.01	58.59	65.41	71.62	54.21	61.71
DANER^{+A+L}	83.77	72.28	77.61	78.65	50.69	61.65	84.93	63.33	72.56	82.44	60.77	69.97
DANER^{+A+S}	77.76	79.35	78.55	74.39	57.16	64.65	80.00	73.07	76.38	77.51	68.55	72.76
DANER$^{+D^*+L}$	**91.99**	94.27	**93.12**	**93.09**	76.36	83.90	**94.06**	87.14	90.50	**94.15**	83.31	88.40
DANER$^{+D^*+S}$	89.22	**96.87**	92.94	89.95	**85.13**	**87.48**	89.06	**95.81**	**92.31**	91.11	**91.73**	**91.42**

It is worth noting that the proposed domain adversarial architecture introduces negligible computational overhead. During the training phase, the additional computations are strictly limited to the parameter-free GRL and the shallow two-layer MLP of the domain classifier (G_d), which add minimal training time per epoch compared to the core Bi-LSTM or BERT feature extractors. Furthermore, the GRL and domain classifier are exclusively utilized during training and completely detached during testing. Consequently, the inference process relies solely on the optimized feature extractor and sequence predictor (G_f and G_y), ensuring that DANER introduces absolutely zero additional inference cost or latency compared to standard baseline models.

4.3. Ablation Study

To rigorously isolate the contribution of the proposed domain adversarial network, the baseline comparisons in Table 2 effectively serve as a comprehensive ablation study. Specifically, the preceding rows represent the ablated baselines (base models without adversarial components), while rows 8–13 detail the performance after incorporating the domain adversarial module (comprising the Gradient Reversal Layer and the Domain Classifier). To systematically analyze these ablation results, we evaluate the performance gains across three levels of underlying feature representations. First, by adding the domain adversarial module to standard baselines lacking massive pre-training (e.g., LSTM-CRF, ID-CNN-CRF), we observe substantial absolute improvements ranging from 5 to over 15 percentage points in F-score, indicating that explicitly enforcing domain invariance is critical for models with limited semantic capacity. Second, when the adversarial module is integrated into models using contextualized Flair embeddings, the resulting complete architectures (DANER^{+A+L} and DANER^{+A+S}) further bridge the cross-

domain gap with consistent gains in recall and overall accuracy. Finally, appending the domain adversarial module to highly optimized pre-trained language models yields the peak F-score of 91.42% (DANER^{+D+S}). Although the absolute F-score gains over ablated BERT-based models and GPT-NER are more modest (approximately 0.5 to 1.5 percentage points), this ablation confirms that adding our domain-adversarial mechanism provides orthogonal benefits, effectively capturing domain-invariant structural features that even massive pre-training paradigms or LLMs struggle to align across diverse industry sectors.

4.4. Parameter Sensitivity

To evaluate the impact of the trade-off parameter λ on the adversarial loss, we conduct parameter sensitivity experiments, with the results summarized in Table 3. The results indicate a clear performance trajectory: as λ increases, the F-score initially improves, peaks at $\lambda = 1$, and subsequently declines. This trend highlights the delicate trade-off governed by λ during adversarial training. When λ is relatively small (e.g., 0.1), the penalty enforced by the gradient reversal layer (GRL) is insufficient. Consequently, the Bi-LSTM feature extractor retains strong domain biases, leading to suboptimal cross-domain generalization. Conversely, when λ is excessively large (e.g., 2.0), the network over-prioritizes misleading the domain classifier. This severe over-penalization degrades the underlying semantic discriminability essential for the primary NER sequence labeling task, inevitably causing a drop in accuracy. Therefore, we recommend $\lambda = 1$ as the optimal equilibrium, as it most effectively balances entity recognition capacity with cross-domain feature invariance.

Table 3. Parameter sensitivity experiments under different λ values.

Methods	All (λ=0.1)			All (λ=0.5)			All (λ=1)			All (λ=2)		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
DANER^{+L}	77.18	40.23	52.89	78.52	42.81	55.41	80.30	44.02	56.86	78.03	41.53	54.21
DANER^{+S}	68.92	51.18	58.74	69.55	53.06	60.20	71.62	54.21	61.71	69.22	52.98	60.02
DANER^{+A+L}	78.31	55.69	65.09	80.91	58.83	68.13	82.44	60.77	69.97	80.71	58.59	67.89
DANER^{+A+S}	73.76	64.91	68.88	75.85	66.82	71.05	77.51	68.55	72.76	75.38	66.19	70.49
DANER$^{+D^*+L}$	90.08	78.95	84.15	92.37	81.09	86.36	94.15	83.31	88.40	92.10	80.86	86.11
DANER$^{+D^*+S}$	87.09	87.31	87.20	89.87	89.65	89.76	91.11	91.73	91.42	89.60	89.35	89.47

4.5. Transferring Performance

In this experiment, we first train baseline model [8]$^{*+L}$, and our proposed model DANER$^{+D^*+L}$ on FIND-2026 datasets. Then, we utilize trained models to detect named entities on MSRA [46] and Weibo [47] datasets to validate the effectiveness of learned domain-invariant features in NER. MSRA and Weibo are popular datasets in Chinese NER, while MSRA is in the news domain and the Weibo NER dataset is drawn from the social media website Sina Weibo (https://www.weibo.com, accessed on 1 April 2026). Table 4 reports the accuracy of the baseline model and the improved model, as well as the improvement in accuracy. In the MSRA dataset, our proposed DANER achieves average absolute improvements of 16.9, 27.4, and 22.4 percentage points in precision (P), recall (R), and F-score (F), respectively. The superior performance is more significant in the Weibo dataset, with an absolute improvement of approximately 28.3 percentage points in F-score. The effectiveness of domain adversarial is strongly proven and confirms our intuitions.

Table 4. We train models (with and without a domain adversarial network) with FIND-2026 datasets and transfer them to two popular datasets (MSRA and Weibo). We report accuracy improvements by our proposed DANER. Precision (P), recall (R), and F-score (F) are reported as percentages. Values in parentheses indicate absolute improvements over the baseline in percentage points.

Datasets	MSRA			Weibo		
	P (%)	**R (%)**	**F (%)**	**P (%)**	**R (%)**	**F (%)**
Baseline	61.3	52.5	56.6	66.8	47.5	55.5
1	78.7 (17.4)	80.0 (27.5)	79.2 (22.6)	81.2 (14.4)	82.0 (34.5)	84.4 (28.9)
2	77.7 (16.4)	81.0 (28.5)	79.7 (23.1)	79.2 (12.4)	83.0 (35.5)	83.9 (28.4)
3	78.3 (17.0)	80.4 (27.9)	79.0 (22.4)	79.7 (12.9)	80.3 (32.8)	83.0 (27.5)
4	78.4 (17.1)	78.6 (26.1)	78.5 (21.9)	79.9 (13.1)	79.0 (31.5)	83.4 (27.9)
5	78.1 (16.8)	79.5 (27.0)	78.7 (22.1)	80.2 (13.4)	81.3 (33.8)	84.2 (28.7)
Avg	78.2 (16.9)	79.9 (27.4)	79.0 (22.4)	80.0 (13.2)	81.1 (33.6)	83.8 (28.3)

4.6. Visualization

To intuitively understand the effect of domain adaptation, we visualize the distribution of the extracted features (where different colors represent different domains). Figure 3 presents t-SNE [48,49] visualizations of the hidden layer features. Figure 3a illustrates the baseline model without domain adaptation, where entities from different sectors are widely separated in the feature space. For instance, the light blue points (industrial) and pink points (financial) are distinctly clustered apart. In contrast, Figure 3b demonstrates that incorporating the DANER procedure effectively mitigates this domain discrepancy. The adversarial training enforces the feature representations to become domain-invariant, thereby forcing the previously distant clusters to heavily overlap and align.

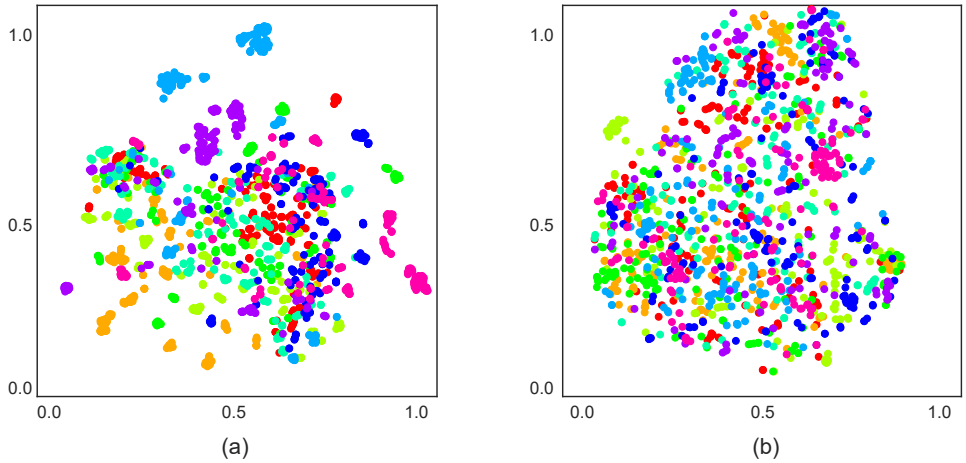


Figure 3. We visualize the learned representations of sampled 300 entities from each sector by t-SNE; they are marked in different colors. (**a**) Baseline without domain adaptation, as we can see, entities of different sectors are generally distributed separately. (**b**) Representations learned by DANER. We observe that feature distributions across domains become more similar. These visualization results verify the effectiveness of domain adversarial networks.

4.7. Case Studies

Table 5 shows case studies demonstrating the effectiveness of the model with and without our proposed DANER. In the first example (E1), there is much ambiguity around the named entity "安琪酵母 (Angel yeast)", which is an abbreviation of Angel yeast Co., Ltd. Baselines without DANER fail to recognize this entity because it is similar to a product name of yeasts. Our proposed methods successfully recognize this abbreviation as an organization. In the second example (E2), both methods detect "佛慈制药 (Foci Pharmaceutical Co., Ltd, Lanzhou, China)". However, the baseline method incorrectly identifies "甘肃药企 (Gansu pharmaceutical enterprises, Lanzhou, China)" as an organization. Thanks to the domain-invariant features, our model avoids this context confusion and detects the true organization name correctly.

Table 5. Case studies evaluating the baseline model with and without our proposed DANER. We utilize $[8]^{*+L}$ as the strong baseline. Red, orange, and green highlight incorrect, missing, and correct entity predictions, respectively.

E1:truth	[安琪酵母]$_{\text{org}}$推出了冷冻面点生坯技术。 [Angel yeast]$_{\text{org}}$ introduced the technology of frozen pastry billet.
E1:w/o daner	安琪酵母推出了冷冻面点生坯技术。 Angel yeast introduced the technology of frozen pastry billet.
E1:w daner	[安琪酵母]$_{\text{org}}$推出了冷冻面点生坯技术。 [Angel yeast]$_{\text{org}}$ introduced the technology of frozen pastry billet.
E2:truth	[佛慈制药]$_{\text{org}}$是甘肃药企走向国际的一个响亮招牌。 [Foci Pharmaceutical]$_{\text{org}}$ is a sign for Gansu pharmaceutical enterprises to go international.
E2:w/o daner	[佛慈制药]$_{\text{org}}$是[甘肃药企]$_{\text{org}}$走向国际的一个响亮招牌。 [Foci Pharmaceutical]$_{\text{org}}$ is a sign for [Gansu pharmaceutical enterprises]$_{\text{org}}$ to go international.
E2:w daner	[佛慈制药]$_{\text{org}}$是甘肃药企走向国际的一个响亮招牌。 [Foci Pharmaceutical]$_{\text{org}}$ is a sign for Gansu pharmaceutical enterprises to go international.

4.8. Error Analysis and Limitations

To better understand the behavior and boundaries of our proposed DANER model, we conducted a detailed error analysis on the cross-domain predictions.

Our analysis indicates that Organization (ORG) entities experience the most significant improvements. In cross-domain scenarios, organizations are frequently represented by ambiguous abbreviations (e.g., "Angel yeast" in Table 5) or act as complex modifiers. Baseline models typically misclassify these as generic nouns or product names due to context shifting. DANER effectively resolves this by mapping context-based token representations into a domain-invariant space, allowing the sequence predictor to recognize the underlying semantic structure of organizations regardless of sector-specific vocabulary.

Despite the overall performance gains, DANER still encounters challenges in two typical scenarios. When evaluating on target domains that feature structurally complex names (e.g., lengthy institutional departments or specific medical terminology), the model occasionally struggles with exact boundary detection. While the domain adversarial layer aligns the feature space, the CRF layer may still prematurely terminate the entity span if the syntactic dependency is overly long. In extreme zero-shot settings where the target domain introduces entirely unseen Out-of-Vocabulary (OOV) entities embedded within grammatical structures absent from the source training data, the model tends to produce False Negatives. Without sufficient invariant contextual cues to anchor the adversarial alignment, the feature extractor fails to trigger the entity recognition mechanism.

Addressing these long-span dependencies and extreme OOV scenarios via structure-aware attention mechanisms remains a promising direction for our future work.

5. Conclusions

Low-resource cross-domain Named Entity Recognition (NER) remains a critical yet challenging problem in natural language processing. In this paper, we addressed this challenge by proposing DANER, an end-to-end domain adversarial neural network designed to explicitly mitigate token-level context shifts. Rather than relying on target-domain data, our sequence tagging architecture integrates a domain classifier that discriminates among multiple source domains, enabling robust zero-shot knowledge transfer. Furthermore, we constructed and introduced FIND-2026, the first large-scale Chinese NER corpus explicitly enriched with fine-grained industry sector annotations. Extensive experiments demonstrate the superiority of our approach: DANER yields substantial absolute improvements over standard baselines on FIND-2026, while providing consistent orthogonal gains to strong pre-trained language models. Finally, the learned domain-invariant features exhibit remarkable zero-shot generalization, yielding absolute F-score improvements of over 22 percentage points on completely unseen public datasets (MSRA and Weibo).

Author Contributions: Z.W.: conceptualization, methodology, software, formal analysis, writing original draft preparation; P.Z.: validation, investigation, conceptualization, project administration, writing review and editing; Y.C.: data curation, resources, visualization, software; Q.L.: funding acquisition, supervision, writing review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the corresponding author.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The FIND-2026 dataset is available from the corresponding author upon reasonable request. The public MSRA and Weibo datasets are described in Refs. [46,47], respectively.

Conflicts of Interest: The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies: During the preparation of this work, the authors used large language models (LLMs) solely to assist with language polishing of the manuscript text. The authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Zhu, P.; Cheng, D.; Yang, F.; et al. Improving Chinese Named Entity Recognition by Large-Scale Syntactic Dependency Graph. *IEEE ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 979–991. https://doi.org/10.1109/taslp.2022.3153261.
2. Feng, X.; Feng, X.; Qin, B.; et al. Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge

Transfer. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 4071–4077. https://doi.org/10.24963/ijcai.2018/566.

3. Chiu, J.P.C.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. https://doi.org/10.1162/tacl_a_00104.
4. Zhu, P.; Li, Y.; Hu, Y.; et al. MCI-GRU: Stock prediction model based on multi-head cross-attention and improved GRU. *Neurocomputing* **2025**, *638*, 130168. https://doi.org/10.1016/j.neucom.2025.130168.
5. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991.
6. Mikolov, T.; Sutskever, I.; Chen, K.; et al. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
7. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
8. Devlin, J.; Chang, M.W.; Lee, K.; et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
9. Sasikumar, N.; Mantri, K.S.I. Transfer learning for low-resource clinical named entity recognition. In Proceedings of the Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 14 July 2023; pp. 514–518.
10. Ganin, Y.; Ustinova, E.; Ajakan, H.; et al. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**. 2096–2030.
11. Bousmalis, K.; Silberman, N.; Dohan, D.; et al. Unsupervised pixel-level domain adaptation with generative adversarial networks. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3722–3731.
12. Rozantsev, A.; Salzmann, M.; Fua, P. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 801–814. https://doi.org/10.1109/tpami.2018.2814042.
13. Lotte, F.; Bougrain, L.; Cichocki, A.; et al. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *J. Neural Eng.* **2018**, *15*, 031005. https://doi.org/10.1088/1741-2552/aab2f2.
14. Wu, D.; Lee, W.S.; Ye, N.; et al. Domain adaptive bootstrapping for named entity recognition. In Proceedings of the 2009 Conference, Singapore, 6–7 August 2009; pp. 1523–1532. https://doi.org/10.3115/1699648.1699699.
15. Chiticariu, L.; Krishnamurthy, R.; Li, Y.; et al. Domain adaptation of rule-based annotators for named-entity recognition tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 1002–1012.
16. Curran, J.R.; Murphy, T.; Scholz, B. Minimising semantic drift with mutual exclusion bootstrapping. In Proceedings of the Conference of the Pacific Association for Computational Linguistics, Melbourne, Australia, 19–21 September 2007; pp. 172–180.
17. Li, Z.; He, Y.; Gu, B.; et al. Diagnosing and Minimizing Semantic Drift in Iterative Bootstrapping Extraction. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 852–865. https://doi.org/10.1109/tkde.2017.2782697.
18. Lafferty, J.; McCallum, A.; Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
19. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. https://doi.org/10.1038/nature14539.
20. Strubell, E.; Verga, P.; Belanger, D.; et al. Fast and accurate entity recognition with iterated dilated convolutions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2670–2680.
21. Lample, G.; Ballesteros, M.; Subramanian, S.; et al. Neural architectures for named entity recognition. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 260–270.
22. Yang, Z.; Salakhutdinov, R.; Cohen, W.W. Transfer learning for sequence tagging with hierarchical recurrent networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–10.
23. Peters, M.E.; Ammar, W.; Bhagavatula, C.; et al. Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1756–1765.
24. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual string embeddings for sequence labeling. In Proceedings of the International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
25. Zhao, D.; Chen, X.; Chen, Y. Named Entity Recognition for Chinese Texts on Marine Coral Reef Ecosystems Based on the BERT-BiGRU-Att-CRF Model. *Appl. Sci.* **2024**, *14*, 5743. https://doi.org/10.3390/app14135743.
26. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. https://doi.org/10.1109/tkde.2009.191.
27. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; et al. Domain Adaptation via Transfer Component Analysis. *IEEE Trans. Neural Netw.*

2011, *22*, 199–210. https://doi.org/10.1109/tnn.2010.2091281.

28. Fernando, B.; Habrard, A.; Sebban, M.; et al. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 2960–2967. https://doi.org/10.1109/iccv.2013.368.
29. Long, M.; Zhu, H.; Wang, J.; et al. Unsupervised domain adaptation with residual transfer networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 136–144.
30. Tzeng, E.; Hoffman, J.; Saenko, K.; et al. Adversarial discriminative domain adaptation. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
31. Chen, Y.H.; Chen, W.Y.; Chen, Y.T.; et al. No more discrimination: Cross city adaptation of road scene segmenters. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1992–2001.
32. Panareda Busto, P.; Iqbal, A.; Gall, J. Open Set Domain Adaptation for Image and Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 413–429. https://doi.org/10.1109/tpami.2018.2880750.
33. Polyak, A.; Taigman, Y.; Wolf, L. Unsupervised Generation of Free-Form and Parameterized Avatars. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 444–459. https://doi.org/10.1109/tpami.2018.2863282.
34. Panareda Busto, P.; Gall, J. Open set domain adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 754–763.
35. Sankaranarayanan, S.; Balaji, Y.; Castillo, C.D.; et al. Generate to adapt: Aligning domains using generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8503–8512.
36. Cao, Z.; Long, M.; Wang, J.; et al. Partial transfer learning with selective adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2724–2732.
37. Zheng, Z.; Teng, L.; Zhang, W.; et al. Knowledge Transfer Learning via Dual Density Sampling for Resource-Limited Domain Adaptation. *IEEE CAA J. Autom. Sin.* **2023**, *10*, 2269–2291. https://doi.org/10.1109/jas.2023.123342.
38. Lu, H.; Wu, J.; Ruan, Y.; et al. A multi-source transfer learning model based on LSTM and domain adaptation for building energy prediction. *Int. J. Electr. Power Energy Syst.* **2023**, *149*, 109024. https://doi.org/10.1016/j.ijepes.2023.109024.
39. Liu, L.; Shang, J.; Ren, X.; et al. Empower sequence labeling with task-aware neural language model. In Proceedings of the Association for the Advancement of Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 1–8.
40. Tzeng, E.; Hoffman, J.; Zhang, N.; et al. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv* **2014**, arXiv:1412.3474.
41. Long, M.; Cao, Y.; Wang, J.; et al. Learning transferable features with deep adaptation networks. In Proceedings of the International Conference on International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 97–105.
42. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1180–1189.
43. Kingma, D.P.; Ba, J.L. Adam: Amethod for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
44. Fan, J.; Furger, A.; Xiu, D. Incorporating Global Industrial Classification Standard Into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator With High-Frequency Data. *J. Bus. Econ. Stat.* **2016**, *34*, 489–503. https://doi.org/10.1080/07350015.2015.1052458.
45. Wang, S.; Sun, X.; Li, X.; et al. GPT-NER: Named entity recognition via large language models. *arXiv* **2023**, arXiv:2304.10428.
46. Levow, G.A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In Proceedings of the SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 108–117.
47. Peng, N.; Dredze, M. Named entity recognition for chinese social media with jointly trained embeddings. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 548–554.
48. van der Maaten, L.; Hinton, G.E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
49. Maaten, L.V.D. Barnes-hut-sne. In Proceedings of the International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2013; pp. 1–11.