Co-Guiding for Multi-Intent Spoken Language Understanding

Bowen Xing and Ivor W. Tsang, *Fellow, IEEE*

***Abstract*—Recent graph-based models for multi-intent SLU have obtained promising results through modeling the guidance from the prediction of intents to the decoding of slot filling. However, existing methods (1) only model the *unidirectional guidance* from intent to slot, while there are bidirectional inter-correlations between intent and slot; (2) adopt *homogeneous graphs* to model the interactions between the slot semantics nodes and intent label nodes, which limit the performance. In this paper, we propose a novel model termed Co-guiding Net, which implements a two-stage framework achieving the *mutual guidances* between the two tasks. In the first stage, the initial estimated labels of both tasks are produced, and then they are leveraged in the second stage to model the mutual guidances. Specifically, we propose two *heterogeneous graph attention networks* working on the proposed two *heterogeneous semantics-label graphs*, which effectively represent the relations among the semantics nodes and label nodes. Besides, we further propose Co-guiding-SCL Net, which exploits the single-task and dual-task semantics contrastive relations. For the first stage, we propose single-task supervised contrastive learning, and for the second stage, we propose co-guiding supervised contrastive learning, which considers the two tasks' mutual guidances in the contrastive learning procedure. Experiment results on multi-intent SLU show that our model outperforms existing models by a large margin, obtaining a relative improvement of 21.3% over the previous best model on MixATIS dataset in overall accuracy. We also evaluate our model on the zero-shot cross-lingual scenario and the results show that our model can relatively improve the state-of-the-art model by 33.5% on average in terms of overall accuracy for the total 9 languages.**

***Index Terms*—Dialog system, graph neural network, multi-task learning, spoken language understanding.**

I. Introduction

SPOKEN language understanding (SLU) [1] is a fundamental task in dialog systems. Its objective is to capture the comprehensive semantics of user utterances, and it typically includes two subtasks: intent detection and slot filling [2]. Intent detection aims to predict the intention of the user utterance and slot filling aims to extract additional information or constraints expressed in the utterance.

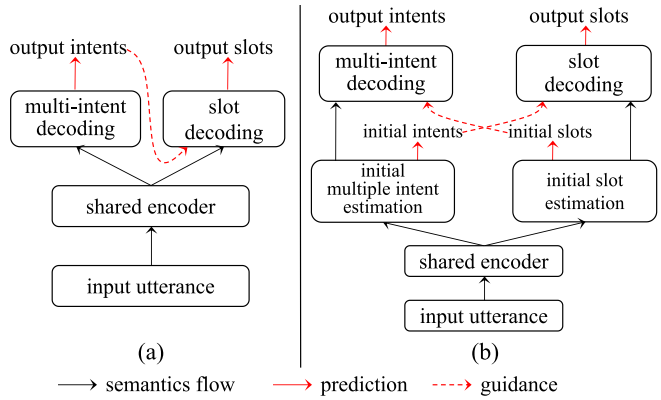


Fig. 1. (a) Previous framework which only models the unidirectional guidance from multi-intent predictions to slot filling. (b) Our framework which models the mutual guidances between the two tasks.

Recently, researchers discovered that these two tasks are closely tied, and a bunch of models [3], [4], [5], [6], [7] are proposed to combine the single-intent detection and slot filling in multi-task frameworks to leverage their correlations.

However, in real-world scenarios, a user usually expresses multiple intents in a single utterance. To this end, [8] begin to tackle the multi-intent detection task and [9] make the first attempt to jointly model the multiple intent detection and slot filling in a multi-task framework. [10] propose an AGIF model to adaptively integrate the fine-grained multi-intent prediction information into the autoregressive decoding process of slot filling via graph attention network (GAT) [11]. And [12] further propose a non-autoregressive GAT-based model which enhances the interactions between the predicted multiple intents and the slot hidden states, obtaining state-of-the-art results and significant speedup.

Despite the promising progress that existing multi-intent SLU joint models have achieved, we discover that they suffer from two main issues:

(1) *Ignoring the guidance from slot to intent:* Since previous researchers realized that "slot labels could depend on the intent" [9], existing models leverage the information of the predicted intents to guide slot filling, as shown in Fig. 1(a). However, they ignore that slot labels can also guide the multi-intent detection

Manuscript received 26 June 2023; revised 30 October 2023; accepted 18 November 2023. Date of publication 29 November 2023; date of current version 3 April 2024. This work was supported by Australian Research Council under Grant DP200101328. The work of Bowen Xing and Ivor W. Tsang was supported by A * STAR Centre for Frontier AI Research. Recommended for acceptance by M. Choudhury. *(Corresponding author: Bowen Xing.)*

Bowen Xing is with the Beijing Key Laboratory of Knowledge Engineering for Materials Science, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: bwxing714@gmail.com).

Ivor W. Tsang is with the CFAR, Agency for Science, Technology and Research, Singapore 138632, also with the IHPC, Agency for Science, Technology and Research, Singapore 138632, also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, and also with the AAII, University of Technology, Ultimo, NSW 2007, Australia (e-mail: ivor_tsang@ihpc.a-star.edu.sg).

Digital Object Identifier 10.1109/TPAMI.2023.3336709

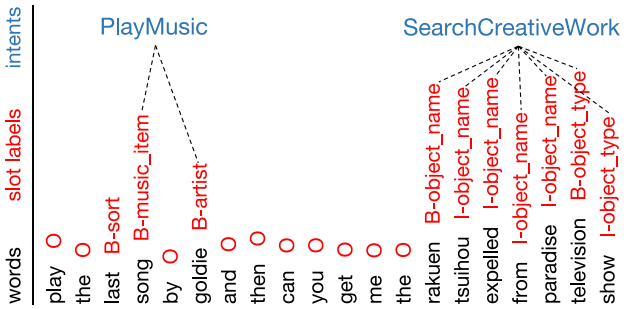


Fig. 2. Illustration of the bidirectional interrelations between intent (blue) and slot (red) labels. The sample is retrieved from MixSNIPS dataset.

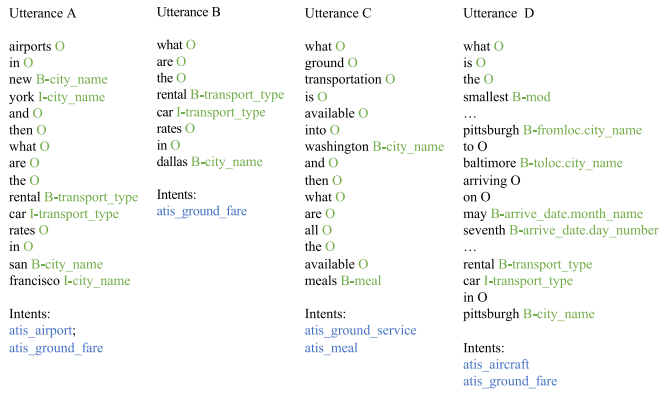


Fig. 3. Illustration of some utterances and their intent and slot labels. Intent labels are in blue while slot labels are in green.

task. Based on our observations, multi-intent detection and slot filling are bidirectionally interrelated and can mutually guide each other. For example, in Fig. 2, not only the intents can indicate the slots, but also the slots can infer the intents. However, in previous works, the only guidance that the multiple intent detection task can get from the joint model is sharing the basic semantics with the slot filling task. As a result, the lack of guidance from slot to intent limits multiple intent detection, and so the joint task.

(2) *Node and edge ambiguity in the semantics-label graph:* [10], [12] apply GATs over the constructed graphs to model the interactions among the slot semantics nodes and intent label nodes. However, their graphs are homogeneous, in which all nodes and edges are treated as the same type. For a slot semantics node, the information from intent label nodes and other slot semantics nodes play different roles, while the homogeneous graph cannot discriminate their specific contributions, causing ambiguity. Therefore, the heterogeneous graphs should be designed to represent the relations among the semantic nodes and label nodes to facilitate better interactions.

In this paper, we propose a novel model termed Co-guiding Net to tackle the above two issues. For the first issue, Co-guiding Net implements a two-stage framework as shown in Fig. 1(b). The first stage produces the initial estimated labels for the two tasks and the second stage leverages the estimated labels as prior label information to allow the two tasks mutually guide each other. For the second issue, we propose two heterogeneous semantics-label graphs (HSLGs): (1) a slot-to-intent semantics-label graph (S2I-SLG) that effectively represents the relations among the intent semantics nodes and slot label nodes; (2) an intent-to-slot semantics-label graph (I2S-SLG) that effectively represents the relations among the slot semantics nodes and intent label nodes. Moreover, two heterogeneous graph attention networks (HGATs) are proposed to work on the two proposed graphs for modeling the guidances from slot to intent and intent to slot, respectively.

To further leverage the subtle semantic differences among the two tasks' instances, aka semantics contrastive relations, which are ignored by previous studies, we propose Co-guiding-SCL Net, which is based on Co-guiding Net and introduces the supervised contrastive learning to draw together the semantics with the same/similar labels and push apart the semantics with different labels. In the *first stage*, since the two tasks are performed individually, we propose two specific single-task supervised contrastive learning mechanisms for multiple intent detection and slot filling, respectively. Since multiple intent detection is a multi-label classification task, the relationships among instances are not simple positive/negative samples. To handle the fine-grained correlations among the multi-intent instances, we propose multi-intent supervised contrastive learning that can dynamically assign a fine-grained weight for each instance regarding the similarity of its intents and the anchor's intents. For slot filling, we adopt the conventional single-label multi-class supervised contrastive learning. In the *second stage*, since the mutual guidances between the two tasks are achieved, there exist dual-task semantics contrastive relations. As shown in the examples in Fig. 3, utterance A and utterance D express similar intents with utterance B (they share the intent label atis_ground_fare). However, we can observe that utterance A and utterance B have more similar sentence-level semantics, while utterance D's semantics is more different. Utterances A and B mention transporting in some cities, while utterance D contains information about transferring from where to where, the mode, and the date. This can be reflected in the fact that utterance D has quite different slot labels from utterance A and utterance B. In the same way, although the word 'san' in utterance A, the word 'dallas' in utterance B and the word 'washington' in utterance C correspond to the same slot label B-city_name, 'san' and 'dallas' should have more similar semantics, while 'washington' should have more different semantics. The reason is that 'san' and 'dallas' have more similar contextual semantics than the contextual semantics of 'washington', which can be reflected from the intent labels of utterance A, utterance B and utterance C. Therefore, intent and slot labels subtly impact the semantics of each other's tasks, which are based on sentence-level and word-level semantics, respectively. And this subtle indicative information can be leveraged as a supervision signal to benefit the dual-task mutual guidances via leveraging the above dual-task contrastive relations. Motivated by this, we propose co-guiding supervised contrastive learning to integrate the dual-task correlations in the contrastive learning procedure. The distances among one task's representations are adjusted

regarding not only the own task's contrastive labels but also the guidance from the other task's contrastive labels. Note that all contrastive learning mechanisms only work in the training process.

The initial version of this work [13] was published on EMNLP 2022 as an oral presentation. Its contributions are three-fold:

1) We propose Co-guiding Net, which implements a two-stage framework allowing multiple intent detection and slot filling mutually guide each other. To the best of our knowledge, this is the first attempt to achieve mutual guidances between the two tasks.
2) We propose two heterogeneous semantics-label graphs as appropriate platforms for the dual-task interactions between semantics nodes and label nodes. And we propose two heterogeneous graph attention networks to model the mutual guidances between the two tasks.
3) Experiment results on two public multi-intent SLU datasets show that our Co-guiding Net significantly outperforms previous models, and model analysis further verifies the advantages of our model.

In this paper, we significantly extend our work from the previous version in the following aspects:

1) We propose Co-guiding-SCL Net, which augments Co-guiding Net with supervised contrastive learning mechanisms to further capture the single-task and dual-task semantics contrastive relations among the samples.
2) For the first stage, we propose the single-task supervised contrastive learning mechanism for both tasks. For multiple intent detection, we propose a novel multi-intent supervised contrastive learning mechanism to capture the dynamic and fine-grained correlations between the multi-intent instances.
3) For the second stage, we propose co-guiding supervised contrastive learning, which can capture the fine-trained dual-task semantics contrastive correlations by jointly considering both tasks' labels as the supervision signal to perform supervised contrastive learning for each task.
4) We conduct extensive experiments on the public multi-intent SLU datasets. Except for LSTM, we also evaluate our model on several pre-trained language model (PTLM) encoders. The experimental results show that our model can achieve significant and consistent improvements over stage-of-the-art models. And the model analysis further verifies the advantages of our proposed dual-task supervised contrastive learning mechanisms.
5) We also evaluate our model on the zero-shot cross-lingual multi-intent SLU task, which has never been explored. The experimental results show that our model can significantly improve the existing best-performing model on the average overall accuracy of the total 9 languages.

The remainder of this paper is organized as follows. In Section II, we summarize the related works of Spoken Language Understanding, Graph Neural Networks for NLP and Contrastive Learning for NLP. And the differences between our method and previous studies are highlighted. Section III elaborates on the details of Co-guiding Net. Section IV depicts the proposed supervised contrastive learning mechanisms in Co-guiding-SCL Net. Experimental results are reported and analyzed in Section V. Note that the task definition of zero-shot cross-lingual multiple intent detection and slot filling as well as the experiments on this task are introduced in Section V-I. Finally, the conclusion of this work and some prospective future directions are provided in Section VI.

II. Related Work

A. Spoken Language Understanding

The correlations between intent detection and slot filling have been widely recognized. To leverage them, a group of models [3], [4], [5], [6], [7], [14], [15], [16], [17], [18], [19] are proposed to tackle the joint task of intent detection and slot filling in a multi-task manner. However, the intent detection modules in the above models can only handle the utterances expressing a single intent, which may not be practical in real-world scenarios, where there are usually multi-intent utterances.

To this end, [8] propose a multi-intent SLU model, and [9] propose the first model to jointly model the tasks of multiple intent detection and slot filling via a slot-gate mechanism. Furthermore, as graph neural networks have been widely utilized in various tasks [20], [21], [22], [23], [24], [25], they have been leveraged to model the correlations between intent and slot. [10] propose an adaptive graph-interactive framework to introduce the fine-grained multiple intent information into slot filling achieved by GATs. More recently, [12] propose another GAT-based model, which includes a non-autoregressive slot decoder conducting parallel decoding for slot filling and achieves the state-of-the-art performance.

Our work also tackles the joint task of multiple intent detection and slot filling. Existing methods only model the one-way guidance from multiple intent detection to slot filling. Besides, they adopt homogeneous graphs and vanilla GATs to achieve the interactions between the predicted intents and slot semantics. Different from previous works, we (1) achieve the mutual guidances between the two tasks; (2) propose the heterogeneous semantics-label graphs to represent the dependencies among the semantics and predicted labels; (3) we propose the Heterogeneous Graph Attention Network to model the semantics-label interactions on the heterogeneous semantics-label graphs; (4) we propose a group of supervised contrastive learning mechanisms to further capture the high-level semantic structures and fine-grained dual-task correlations.

B. Graph Neural Networks for NLP

In recent years, graph neural networks have been widely adopted in various NLP tasks. Some works [21], [23], [26], [27] leverage the GAT and GCN to encode syntactic information for target sentiment classification. CGR-Net [28] models the interactions among the emotion-cause pair extraction and the two subtasks through the multi-task relational graph. In dialog understanding, DARER [24], [29] applies the relational graph convolutional network [30] over the constructed speaker-aware temporal graph and dual-task temporal graph to capture the

relational temporal information. In spoken language understanding, AGIF [10] and GL-GIN [12] leverage graph structures and GNNs to model the intent-slot correlation. ReLa-Net [31] adopts a heterogeneous label graph to model the dual-task label dependencies. In this paper, we propose two heterogeneous graphs to provide an appropriate platform for dual-task semantics-label interactions, which is achieved by our proposed heterogeneous attention networks.

C. Contrastive Learning for NLP

Contrastive learning has been leveraged to improve semantic representations in different NLP tasks [32], [33], [34]. In natural language inference, pairwise supervised CL [32] propose to utilize high-level categorical concept encoding to bridge semantic entailment and contradiction understanding. Hierarchy-guided contrastive learning [34] is proposed to directly incorporate the hierarchy into the text encoder for hierarchical text classification. GL-CLEF [35] performs unsupervised contrastive learning to achieve the cross-lingual semantics alignment to improve zero-shot cross-lingual intent detection and slot filling. In this paper, we propose single-task supervised contrastive learning to further capture the single-task semantics contrastive relations in the first stage. And we propose co-guiding supervised contrastive learning to capture the dual-task semantics contrastive relations in the second stage.

III. Co-Guiding

Problem Definition: Given a input utterance denoted as $U = \{u_i\}_1^n$, multiple intent detection can be formulated as a multi-label classification task that outputs multiple intent labels corresponding to the input utterance. And slot filling is a sequence labeling task that maps each u_i into a slot label.

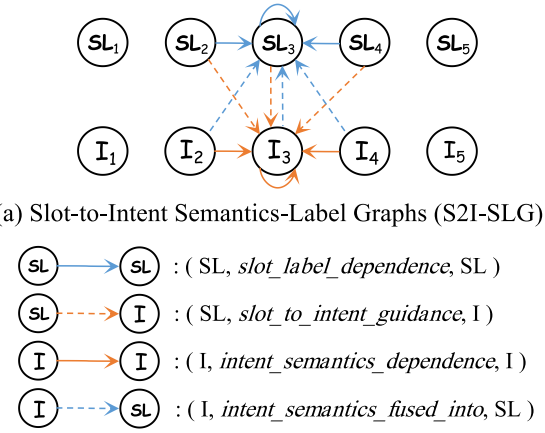
Next, before diving into the details of Co-guiding Net's architecture, we first introduce the construction of the two heterogeneous graphs.

A. Graph Construction

1) Slot-to-Intent Semantics-Label Graph: To provide an appropriate platform for modeling the guidance from the estimated slot labels to multiple intent detection, we design a slot-to-intent semantics-label graph (S2I-SLG), which represents the relations among the semantics of multiple intent detection and the estimated slot labels. S2I-SLG is a heterogeneous graph and an example is shown in Fig. 4(a). It contains two types of nodes: intent semantics nodes[1] (e.g., I_1,..., I_5) and *s*lot *l*abel (SL) nodes (e.g., SL_1,..., SL_5). And there are four types of edges in S2I-SLG, as shown in Fig. 4(b). Each edge type corresponds to an individual kind of information aggregation on the graph.

Mathematically, the S2I-SLG can be denoted as $\mathcal{G}_{s2i} = (\mathcal{V}_{s2i}, \mathcal{E}_{s2i}, \mathcal{A}_{s2i}, \mathcal{R}_{s2i})$, in which $\mathcal{V}_{s2i}$ is the set of all nodes, $\mathcal{E}_{s2i}$ is the set of all edges, $\mathcal{A}_{s2i}$ is the set of two node types and $\mathcal{R}_{s2i}$ is the set of four edge types. Each node v_{s2i} and each edge e_{s2i} are associated with their type mapping functions $\tau(v_{s2i}) : \mathcal{V}_{s2i} \rightarrow \mathcal{A}_{s2i}$ and $\phi(e_{s2i}) : \mathcal{E}_{s2i} \rightarrow \mathcal{R}_{s2i}$. For instance, in Fig. 4, the SL_2 node belongs to $\mathcal{V}_{s2i}$, while its node type SL belongs to $\mathcal{A}_{s2i}$; the edge from SL_2 to I_3 belongs to $\mathcal{E}_{s2i}$, while its edge type *slot_to_intent_guidance* belongs to $\mathcal{R}_{s2i}$. Besides, edges in S2I-SLG are based on local connections. For example, node I_i is connected to $\{\mathrm{I}_{i-w}, \ldots, \mathrm{I}_{i+w}\}$ and $\{\mathrm{SL}_{i-w}, \ldots, \mathrm{SL}_{i+w}\}$, where w is a hyper-parameter of the local window size.

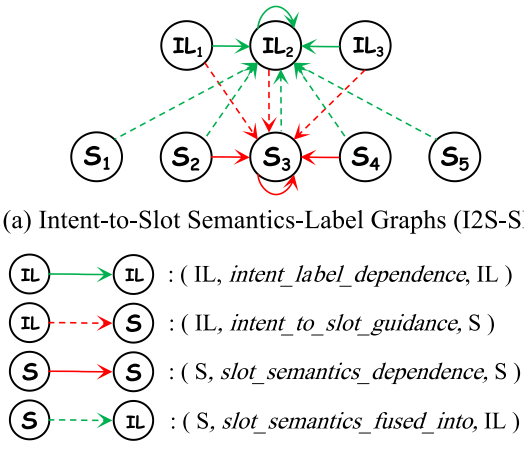
[1]Each word corresponds to a semantics node. In LSTM setting, semantics node representation is each word's hidden state. In PTLM setting, semantics node representation is each word's first token's hidden state.



(a) Slot-to-Intent Semantics-Label Graphs (S2I-SLG)

(b) Relation types of S2I-SLG

Fig. 4. Illustration of S2I-SLG and its relation types. w.l.o.g, only the edges directed into SL_3 and I_3 are shown, and the local window size is 1.



(a) Intent-to-Slot Semantics-Label Graphs (I2S-SLG)

(b) Relation types of I2S-SLG

Fig. 5. Illustration of I2G-SLG and its relation types. w.l.o.g, only the edges directed into IL_3 and S_3 are shown, and the local window size is 1.

2) Intent-to-Slot Semantics-Label Graph: To present a platform for accommodating the guidance from the estimated intent labels to slot filling, we design an intent-to-slot semantics-label graph (I2S-SLG) that represents the relations among the slot semantics nodes and the intent label nodes. I2S-SLG is also a heterogeneous graph and an example is shown in Fig. 5(a). It contains two types of nodes: slot semantics nodes (e.g., S_1,..., S_5) and *i*ntent *l*abel (IL) nodes (e.g., IL_1,..., IL_5). And Fig. 5(b)

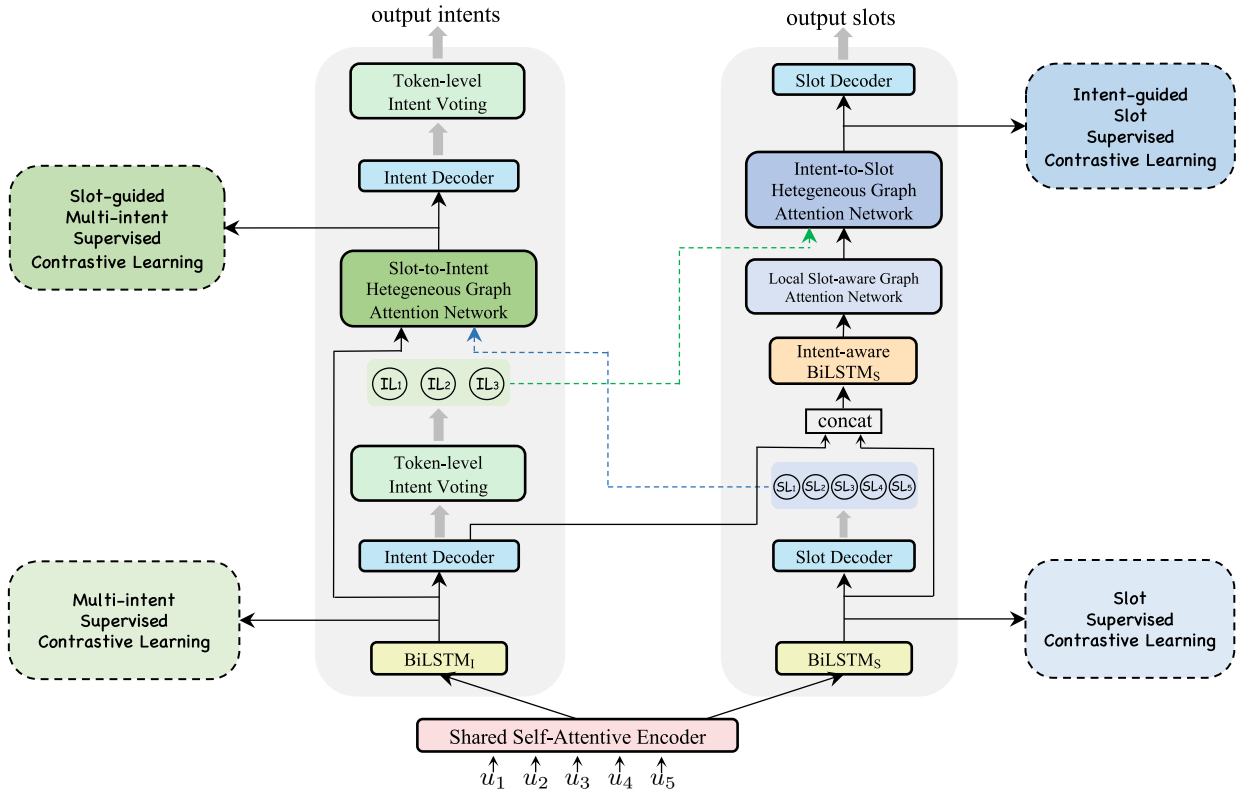


Fig. 6. Architecture of Co-guiding Net and Co-guiding-SCL Net. Dashed boxes denote the contrastive learning modules included in Co-guiding-SCL Net while not in Co-guiding Net. Dashed lines denote that the contrastive learning modules only work in the training procedure. Each HGAT is triggered by its own task's semantics and the counterpart's predicted labels. The green and blue dashed arrow lines denote the projected label representations from the predicted intents and slots, respectively. The green solid arrow line denotes the intent distribution generated by the Intent Decoder at the first stage.

shows the four edge types. Each edge type corresponds to an individual kind of information aggregation on the graph.

Mathematically, the I2S-SLG can be denoted as $\mathcal{G}_{i2s} = (\mathcal{V}_{i2s}, \mathcal{E}_{i2s}, \mathcal{A}_{i2s}, \mathcal{R}_{i2s})$. Each node v_{i2s} and each edge e_{i2s} are associated with their type mapping functions $\tau(v_{i2s})$ and $\phi(e_{i2s})$. The connections in I2S-SLG are a little different from S2I-SLG. Since intents are sentence-level, each IL node is globally connected with all nodes. For S_i node, it is connected to $\{\mathrm{S}_{i-w}, \ldots, \mathrm{S}_{i+w}\}$ and $\{\mathrm{IL}_1, \ldots, \mathrm{IL}_m\}$, where w is the local window size and m is the number of estimated intents.

B. Model Architecture

In this section, we introduce the details of our Co-guiding Net, whose architecture is shown in Fig. 6.

1) Shared Self-Attentive Encoder: Following [10], [12], we adopt a shared self-attentive encoder to produce the initial hidden states containing the basic semantics. It includes a BiLSTM and a self-attention module. BiLSTM captures the temporal dependencies:

$$h_i = \mathrm{BiLSTM}\left(x_i, h_{i-1}, h_{i+1}\right) \tag{1}$$

where x_i is the word vector of u_i. Now we obtain the context-sensitive hidden states $\hat{\boldsymbol{H}} = \{\hat{h}_i\}_1^n$.

Self-attention captures the global dependencies:

$$\boldsymbol{H'} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

where $\boldsymbol{H'}$ is the global contextual hidden states output by self-attention; $\boldsymbol{Q}, \boldsymbol{K}$ and $\boldsymbol{V}$ are matrices obtained by applying different linear projections on the input utterance word vector matrix.

Then we concatenate the output of BiLSTM and self-attention to form the output of the shared self-attentive encoder: $\boldsymbol{H} = \hat{\boldsymbol{H}} \| \boldsymbol{H'}$, where $\boldsymbol{H} = \{h_i\}_1^n$ and $\|$ denotes concatenation operation.

2) Initial Estimation. Multiple Intent Detection: To obtain the task-specific features for multiple intent detection, we apply a BiLSTM layer over $\boldsymbol{H}$:

$$h_i^{[I,0]} = \mathrm{BiLSTM_I}\left(h_i, h_{i-1}^{[I,0]}, h_{i+1}^{[I,0]}\right) \tag{3}$$

Following [10], [12], we conduct token-level multi-intent detection. Each $h_i^{[I,0]}$ is fed into the intent decoder. Specifically, the intent label distributions of the i-th word are obtained by:

$$y_i^{[I,0]} = \mathrm{sigmoid}\left(\boldsymbol{W}_I^1\left(\sigma(\boldsymbol{W}_I^2 \boldsymbol{h}_i^{[I,0]} + \boldsymbol{b}_I^2)\right) + \boldsymbol{b}_I^1\right) \tag{4}$$

where σ denotes the non-linear activation function; W_* and b_* are model parameters.

Then the estimated sentence-level intent labels $\{\text{IL}_1, \ldots, \text{IL}_m\}$ are obtained by the token-level intent voting [12].

Slot Filling: [12] propose a non-autoregressive paradigm for slot filling decoding, which achieves significant speedup. In this paper, we also conduct parallel slot filling decoding.

We first apply a BiLSTM over $\boldsymbol{H}$ to obtain the task-specific features for slot filling:

$$h_i^{[S,0]} = \text{BiLSTM}_\text{S}\left(h_i, h_{i-1}^{[S,0]}, h_{i+1}^{[S,0]}\right) \tag{5}$$

Then use a softmax classifier to generate the slot label distribution for each word:

$$y_i^{[S,0]} = \text{softmax}\left(\boldsymbol{W}_S^1\left(\sigma(\boldsymbol{W}_S^2\boldsymbol{h}_i^{[S,0]} + \boldsymbol{b}_S^2)\right) + \boldsymbol{b}_S^1\right) \tag{6}$$

And the estimated slot label for each word is obtained by $\text{SL}_i = \arg\max(y_i^{[S,0]})$.

3) Heterogeneous Graph Attention Network: State-of-the-art models [10], [12] use a homogeneous graph to connect the semantic nodes of slot filling and the intent label nodes. And GAT [11] is adopted to achieve information aggregation. In Section I, we propose that this manner cannot effectively learn the interactions between one task's semantics and the estimated labels of the other task. To tackle this issue, we propose two heterogeneous graphs (S2I-SLG and I2S-SLG) to effectively represent the relations among the semantic nodes and label nodes. To model the interactions between semantics and labels on the proposed graphs, we propose a Heterogeneous Graph Attention Network (HGAT). When aggregating the information into a node, HGAT can discriminate the specific information from different types of nodes along different relations. And two HGATs (S2I-HGAT and I2S-HGAT) are applied on S2I-SLG and I2S-SLG, respectively. Specifically, S2I-HGAT can be formulated as follows:

$$\begin{aligned} h_i^{l+1} &= \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_{s2i}^i} W_{s2i}^{[r,k,1]}\alpha_{ij}^{[r,k]}h_j^l\right), r = \phi\left(e_{s2i}^{[j,i]}\right) \\ \alpha_{ij}^{[r,k]} &= \frac{\exp\left(\left(W_{s2i}^{[r,k,2]}h_i^l\right)\left(W_{s2i}^{[r,k,3]}h_j^l\right)^\top/\sqrt{d}\right)}{\sum_{u\in\mathcal{N}_{s2i}^{r,i}}\exp\left(\left(W_{s2i}^{[r,k,2]}h_i^l\right)\left(W_{s2i}^{[r,k,3]}h_u^l\right)^\top/\sqrt{d}\right)} \end{aligned} \tag{7}$$

where K denotes the total head number; $\mathcal{N}_{s2i}^i$ denotes the set of incoming neighbors of node i on S2I-SLG; $W_{s2i}^{[r,k,*]}$ are weight matrices of edge type r on the k-th head; $e_{s2i}^{[j,i]}$ denotes the edge from node j to node i on S2I-SLG; $\mathcal{N}_{s2i}^{r,i}$ denotes the nodes connected to node i with r-type edges on S2I-SLG; d is the dimension of node hidden state.

I2S-HGAT can be derived like (7).

4) Intent Decoding With Slot Guidance: In the first stage, we obtain the initial intent features $H^{[I,0]} = \{h_i^{I,0}\}_i^n$ and the initial estimated slot labels sequence $\{\text{SL}_1, \ldots, \text{SL}_n\}$. Now we project the slot labels into vector form using the slot label embedding matrix, obtaining $E_{sl} = \{e_{sl}^1, \ldots, e_{sl}^n\}$.

Then we feed $H^{[I,0]}$ and E_{sl} into S2I-HGAT to model their interactions, allowing the estimated slot label information to guide the intent decoding:

$$H^{[I,L]} = \text{S2I-HGAT}\left([H^{[I,0]}, E_{sl}], \mathcal{G}_{s2i}, \theta_I\right) \tag{8}$$

where $[H^{[I,0]}, E_{sl}]$ denotes the input node representation; θ_I denotes S2I-HGAT's parameters. L denotes the total layer number.

Finally, $H^{[I,L]}$ is fed to intent decoder, producing the intent label distributions for the utterance words: $Y^{[I,1]} = \{y_i^{[I,1]}, \ldots, y_n^{[I,1]}\}$. And the final output sentence-level intents are obtained via applying token-level intent voting over $Y^{[I,1]}$.

5) Slot Decoding With Intent Guidance: Intent-aware BiLSTM: Since the B-I-O tags of slot labels have temporal dependencies, we use an intent-aware BiLSTM to model the temporal dependencies among slot hidden states with the guidance of estimated intents:

$$\tilde{h}_i^{[S,0]} = \text{BiLSTM}\left(y_i^{[I,0]} \Vert h_i^{[S,0]}, \tilde{h}_{i-1}^{[S,0]}, \tilde{h}_{i+1}^{[S,0]}\right) \tag{9}$$

I2S-HGAT: We first project the estimated intent labels $\{\text{IL}_j\}_1^m$ into vectors using the intent label embedding matrix, obtaining $E_{il} = \{e_{il}^1, \ldots, e_{il}^m\}$. Then we feed $\tilde{H}^S$ and E_{il} into I2S-HGAT to model their interactions, allowing the estimated intent label information to guide the slot decoding:

$$H^{[S,L]} = \text{I2S-HGAT}\left([\tilde{H}^S, E_{il}], \mathcal{G}_{i2s}, \theta_S\right) \tag{10}$$

where $[\tilde{H}^{[S]}, E_{il}]$ denotes the input node representation; θ_S denotes I2S-HGAT's parameters.

Finally, $H^{[S,L]}$ is fed to slot decoder, producing the slot label distributions for each word: $Y^{[S,1]} = \{y_i^{[S,1]}, \ldots, y_n^{[S,1]}\}$. And the final output slot labels are obtained by applying $\arg\max$ over $Y^{[S,1]}$.

C. Training Objective

1) Loss Function: The loss function for multiple intent detection is:

$$\begin{aligned} \text{CE}(\hat{y}, y) &= \hat{y}\log(y) + (1-\hat{y})\log(1-y) \\ \mathcal{L}_I &= \sum_{t=0}^{1}\sum_{i=1}^{n}\sum_{j=1}^{N_I}\text{CE}\left(\hat{y}_i^I[j], y_i^{[I,t]}[j]\right) \end{aligned} \tag{11}$$

And the loss function for slot filling is:

$$\mathcal{L}_S = \sum_{t=0}^{1}\sum_{i=1}^{n}\sum_{j=1}^{N_S}\hat{y}_i^S[j]\log\left(y_i^{[S,t]}[j]\right) \tag{12}$$

where N_I and N_S denote the total numbers of intent labels and slot labels; $\hat{y}_i^I$ and $\hat{y}_i^S$ denote the ground-truth intent labels and slot labels.

2) Margin Penalty: The core of our model is to let the two tasks mutually guide each other. Intuitively, the predictions in the second stage should be better than those in the first stage. To force our model to obey this rule, we design a margin penalty ($\mathcal{L}^{mp}$) for each task, whose aim is to improve the probabilities of the correct labels. Specifically, $\mathcal{L}_I^{mp}$ and $\mathcal{L}_S^{mp}$ are formulated

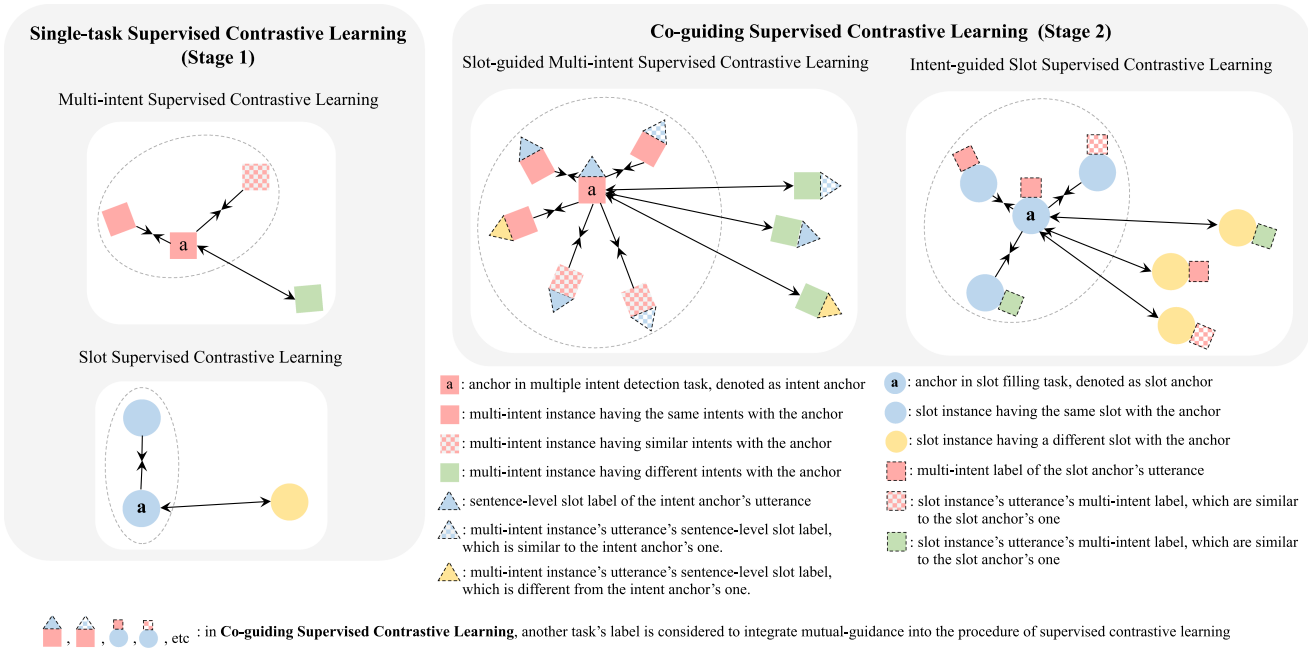


Fig. 7. Conceptual illustration of single-task supervised contrastive learning, which is performed in stage 1, and co-guiding supervised contrastive learning, which is performed in stage 2.

as:

$$\mathcal{L}_I^{mp} = \sum_{i=1}^{n} \sum_{j=1}^{N_I} \hat{y}_i^I[j] \ \max\left(0, y_i^{[I,0]}[j] - y_i^{[I,1]}[j]\right)$$

$$\mathcal{L}_S^{mp} = \sum_{i=1}^{n} \sum_{j=1}^{N_S} \hat{y}_i^S[j] \max\left(0, y_i^{[S,0]}[j] - y_i^{[S,1]}[j]\right) \quad (13)$$

3) Model Training: The training objective $\mathcal{L}$ is the weighted sum of loss functions and margin regularizations of the two tasks:

$$\mathcal{L} = \gamma\left(\mathcal{L}_I + \beta_I \mathcal{L}_I^{mp}\right) + (1-\gamma)\left(\mathcal{L}_S + \beta_S \mathcal{L}_S^{mp}\right) \quad (14)$$

where γ is the coefficient balancing the two tasks; β_I and β_S are the coefficients of the margin regularization for the two tasks.

IV. Co-Guiding-SCL Net

Existing methods suffer from three issues: 1) ignoring the guidance from slot to intent; 2) Node and edge ambiguity in the semantics-label graph; 3) ignoring the subtle semantic differences among the two tasks' instances. Co-guiding Net depicted in Section III is proposed to tackle the first two issues. In this section, we focus on solving the third issue. As stated in Section I, the single-task and dual-task semantics contrastive relations can benefit single-task reasoning and dual-task mutual guidances. Based on Co-guiding Net, we propose Co-guiding-SCL Net, which is augmented with our proposed single-task supervised contrastive learning and Co-guiding supervised contrastive learning, which are illustrated in Fig. 7. The description of the notations in this section can be found in Table I.

Since our model performs supervised contrastive learnings, inspired by [36], we maintain a set of sample queues that store not only the previously encoded features but also their labels: $Q_u^0, Q_s^0, Q_u^1, Q_s^1, Q_l^I, Q_l^S$ and Q_l^{ss}, whose descriptions can be found in Table I. Note that the sentence-level slot label is used to provide sentence-level slot guidance in the proposed slot-guided intent supervised contrastive learning (Section IV-B1). Since it is not given in the datasets, we construct it by ourself and details can be found in Section IV-B1. After the current batch, we update the sample queues with the current batch's features and labels while dequeuing the oldest ones.

Next, we depict our proposed supervised contrastive learning mechanisms.

A. Single-Task Supervised Contrastive Learning

In the first stage, initial estimation is performed to predict the initial labels that provide guidance for the other task. The initial estimation is only based on the semantics of the current task. As we state in Section I, there exist inherent semantics contrastive relations among the representations of each task. Intuitively, the semantics representations corresponding to the same/similar intents or the same slots should be close to each other in the representation space. Contrastively, the semantics representations corresponding to the same/similar intents or the same slots should be near to each other in the representation space. To achieve this, we propose two supervised contrastive learning mechanisms for multiple intent detection and slot filling, respectively.

1) Multi-Intent Supervised Contrastive Learning: The function of this contrastive learning mechanism is to pull together the utterance representations that have the same/similar intent labels, while pushing apart the ones having different intent labels. The anchor is $h_u^{[I,0]}$ and the contrastive instances are from Q_u^0. Unlike the conventional single-label multi-class

TABLE I
DESCRIPTIONS OF NOTATIONS USED IN SECTION IV

Notation	Description
$h_u^{[I,0]}$	Current sample's utterance representation in stage 1. $h_u^{[I,0]} = \frac{1}{n}\sum_i^n h_i^{[I,0]}$
$h_i^{[S,0]}$	Current sample's i-th word representation of slot filling task in stage 1.
$h_u^{[I,1]}$	Current sample's utterance representation in stage 2. $h_u^{[I,1]} = \frac{1}{n}\sum_i^n h_i^{[I,1]}$
$h_i^{[S,1]}$	Current sample's i-th word representation of slot filling task in stage 2.
l_I	Current sample's intent label vector
l_S^j	Current sample's j-th word's slot label vector
l^{ss}	Current sample's sentence-level slot label vector
l^J	Current sample's joint-task label vector for slot-guided multi-intent supervised contrastive learning.
$l^{J'}$	Current sample's joint-task label vector for intent-guided slot supervised contrastive learning.
K	The size of sample queues
$Q_u^0 = \{h_{[uq,0]}^k\}_K^k$	Sample queue of the utterance representations in stage 1.
$h_{[uq,0]}^k$	k-th instance in Q_u^0
$Q_s^0 = \{h_{[sq,0]}^{[k,0]}, ..., h_{[sq,0]}^{[k,j]}, ...h_{[sq,0]}^{[k,n]}\}_K^k$	Sample queue of the word representations of slot filling task in stage 1.
$h_{[sq,0]}^{[k,j]}$	k-th instance's j-th word representations in Q_s^0
$Q_u^1 = \{h_{[uq,1]}^k\}_K^k$	Sample queue of the utterance representations in stage 2.
$h_{[uq,1]}^k$	k-th instance in Q_u^1
$Q_s^1 = \{h_{[sq,1]}^{[k,0]}, ..., h_{[sq,1]}^{[k,j]}, ...h_{[sq,1]}^{[k,n]}\}_K^k$	Sample queue of the word representations of slot filling task in stage 2.
$h_{[sq,1]}^{[k,j]}$	k-th instance's j-th word representations in Q_s^1
$Q_l^I = \{l_k^I\}_K^k$	Sample queue of multi-hot intent label vectors
l_k^I	k-th instance's intent label vector
$Q_l^S = \{l_S^{[k,0]}, ..., l_S^{[k,j]}, ..., l_S^{[k,n]}\}_K^k$	Sample queue of one-hot slot label vectors
$l_S^{[k,j]}$	The one-hot slot label vector of the k-th instance's j-th word
$Q_l^s s = \{l_k^{ss}\}_K^k$	Sample queue of sentence-level slot label vectors
l_k^{ss}	k-th instance's sentence-level slot label vectors
l_k^J	k-th instance's joint-task label vector for slot-guided multi-intent supervised contrastive learning.
$l_k^{J'}$	k-th instance's joint-task label vector for intent-guided slot supervised contrastive learning.

TABLE II
HYPER-PARAMETERS TUNED IN OUR EXPERIMENTS

word embedding dimension	100, 128, 200, 256, 300
label embedding dimension	100, 128, 200, 256, 300
hidden state dimension	100, 128, 200, 256, 300
layer number of GNNs	2,3,4
learning rate	5e-4, 1e-3, 5e-3
weight decay	0, 1e-6
β_I	1e-8, 1e-6, 1e-4, 1e-2, 1, 10
β_S	1e-8, 1e-6, 1e-4, 1e-2, 1, 10
τ	0.05, 0.07, 0.1
η_I	1e-4, 1e-3, 0.01, 0.1, 1
η_S	1e-4, 1e-3, 0.01, 0.1, 1

supervised contrastive learning, our proposed multi-intent supervised contrastive learning can handle the fine-grained and dynamic relations among the multi-intent instances. Specifically, this contrastive learning mechanism can be formulated as:

$$\mathcal{L}_{\text{SCL}}^{\text{MI}} = -\sum_k^K \mu_k \log \frac{e^{s(h_u^{[I,0]}, h_{[uq,0]}^k)}}{\sum_j^K e^{s(h_u^{[I,0]}, h_{[uq,0]}^j)}}$$

$$\mu_k = \frac{l_I \odot l_I^k}{\sum_j^K l_I \odot l_I^j} \tag{15}$$

where $\odot$ denotes Hadamard product, $s(a,b) = \frac{a^T \cdot b}{\|a\| \cdot \|b\| \cdot \tau}$ is the cosine similarity function, and τ denotes contrastive learning temperature. $l_I \odot l_I^k$ denotes the golden similarity between the anchor and the k-th multi-intent instance. A large $l_I \odot l_I^k$ denotes the k-th instance is quite similar to the anchor, leading to a large μ_k assigned to the loss function to pull them closer. Instead, if they have totally different labels, $l_I \odot l_I^k = 0$ and then $\mu_k = 0$. In this case, $s(h_u^{[I,0]}, h_{[uq,0]}^j)$, which denotes their distance, only appears in the denominator. As a result, the anchor and k-th multi-intent instance will be pushed apart by the negative gradient.

2) Slot Supervised Contrastive Learning: Since each word corresponds to only one slot label, this contrastive learning mechanism is conventional single-label multi-class supervised contrastive learning. It aims to pull together the word representations that correspond to the same slot, while pushing apart the ones corresponding to different slots. The anchor is $h_i^{[S,0]}$ and the contrastive instances are from Q_s^0. Specifically, this contrastive learning mechanism can be formulated as:

$$\mathcal{L}_{\text{SCL}}^{\text{S}} = -\sum_i^n \sum_j^n \sum_k^K \frac{l_S^i \odot l_S^{[k,j]}}{M_i} \log \frac{e^{s(h_i^{[S,0]}, h_{[sq,0]}^{[k,j]})}}{E_i}$$

$$M_i = \sum_j^n \sum_k^K l_S^i \odot l_S^{[k,j]}$$

$$E_i = \sum_j^n \sum_j^K e^{s(h_i^{[S,0]}, h_{[sq,0]}^{[k,j]})} \tag{16}$$

$l_S^i \odot l_S^{[k,j]}$ equals 1 or 0, indicating the j-th word representation of the k-th instance in Q_s^0 is the positive sample or negative sample of the i-th word representation of the current utterance.

B. Co-Guiding Supervised Contrastive Learning

In the second stage, the mutual guidances between the two tasks are achieved. The representations (e.g., $h^{[I,1]_u}, h^{[S,1]_i}$) in the second stage contain two kinds of information: (1) the own task's semantics that can indicate the own task's labels; (2) the other task's initial label information that provides dual-task guidance. Among the semantics representations of the two tasks, there exist dual-task semantics contrastive relations, which have been stated in Section I. Therefore, we propose co-guiding supervised contrastive learning to integrate dual-task correlations into the contrastive learning procedure, which jointly considers both tasks' labels as the supervision signal to perform supervised contrastive learning. Next, we introduce the details of slot-guided multi-intent supervised contrastive learning.

1) Slot-Guided Multi-Intent Supervised Contrastive Learning: Multiple intent detection is a sentence-level classification task. Although slot filling is word-level, the summarization of all of the slot labels in an utterance can provide sentence-level slot semantics. For the utterances having the same intents, some of them may have different sentence-level slot semantics, which can be leveraged to discriminate the representations of these utterances. And for the utterances having different intents, some of them may have similar sentence-level slot semantics, which can be leveraged to adjust the distances among their representations. , learning better representations. Slot-guided multi-intent supervised contrastive learning is proposed to achieve the above two aspects.

First, we have to construct the sentence-level slot label by ourselves because it is not provided in the datasets. The current utterance's sentence-level slot label vector is obtained by: $l^{ss} = \frac{\sum_{i,l_S^i \neq \mathrm{O}}^{n} l_S^i}{\sum_{i=1,l_S^i \neq \mathrm{O}}^{N} 1}$. The value of each dimension in l^{ss} ranges from 0 to 1, which can be regarded as a score reflecting the degree of the corresponding slot for the sentence-level slot semantics. Then we construct the joint-task label by concatenating l_I with weighted l^{ss}: $l_J = \mathrm{concat}(l_I, \lambda^I * l^{ss})$, where λ^I is a hyper-parameter.

Then the formulation of slot-guided multi-intent supervised contrastive learning is:

$$\mathcal{L}_{\mathrm{SCL}}^{\mathrm{SGMI}} = -\sum_{k}^{K} \mu_k \log \frac{e^{s(h_u^{[I,1]}, h_{[uq,1]}^k)}}{\sum_j^K e^{s(h_u^{[I,1]}, h_{[uq,1]}^j)}}$$

$$\mu_k = \frac{l^J \odot l_k^J}{\sum_j^K l^J \odot l_j^J} \tag{17}$$

Note that $l^J \odot l_k^J = l_I \odot l_k^I + \lambda^I * \lambda^I * l^{ss} \odot l_k^{ss}$. In this way, λ^I can control the extent the slot label is integrated for slot-guided multi-intent supervised contrastive learning.

2) Intent-Guided Slot Supervised Contrastive Learning: Generally, the semantics of the intents expressed in an utterance is contained in each word's representation. For some word representations from different utterances, their corresponding utterances may have different intents. Even if they correspond to the same slot, their semantics are somehow different regarding the different intent semantics they contain. And some other words' corresponding utterances may have the same/similar intents. Even if they correspond to different slots, their semantics may be not quite different regarding the same/smilar intent semantics they contain. The above two aspects can be leveraged to further discriminate the word representations corresponding to the same slot and adjust the distance among the ones corresponding to different slots. To this end, we propose intent-guided slot supervised contrastive learning.

First, we construct the joint-task label $l_i^{J'} = \mathrm{concat}(l_S^i, \lambda^S * l_I)$, where λ^S is a hyper-parameter. Then the intent-guided slot supervised contrastive learning can be formulated as:

$$\mathcal{L}_{\mathrm{SCL}}^{\mathrm{IGS}} = -\sum_{i}^{n}\sum_{j}^{n}\sum_{k}^{K} \frac{l^{J'} \odot l_k^{J'}}{M_i} \log \frac{e^{s(h_i^{[S,1]}, h_{[sq,1]}^{[k,j]})}}{E_i}$$

$$M_i = \sum_{j}^{n}\sum_{k}^{K} l_i^{J'} \odot l_{[k,j]}^{J'}$$

$$E_i = \sum_{j}^{n}\sum_{j}^{K} e^{s(h_i^{[S,1]}, h_{[sq,1]}^{[k,j]})} \tag{18}$$

Note that $l_i^{J'} \odot l_{[k,j]}^{J'} = l_S^i \odot l_S^{[k,j]} + \lambda^S * \lambda^S * l_I \odot l_k^I$. In this way, λ^S can control the extent that intent labels are integrated for intent-guided slot supervised contrastive learning.

C. Training Objective

The final loss of Co-guiding-SCL Net is the sum of the loss of Co-guiding Net and all contrastive loss terms:

$$\mathcal{L} = \gamma\left(\mathcal{L}_I + \beta_I \mathcal{L}_I^{mp}\right) + (1-\gamma)\left(\mathcal{L}_S + \beta_S \mathcal{L}_S^{mp}\right) + \eta_I\left(\mathcal{L}_{\mathrm{SCL}}^{\mathrm{MI}} + \mathcal{L}_{\mathrm{SCL}}^{\mathrm{SGMI}}\right) + \eta_S\left(\mathcal{L}_{\mathrm{SCL}}^{\mathrm{S}} + \mathcal{L}_{\mathrm{SCL}}^{\mathrm{IGS}}\right) \tag{19}$$

where η_I and η_S are hyper-parameters that balance the contrastive loss terms. Note that all contrastive learning mechanisms only participate in the training process. Co-guiding Net and Co-guiding-SCL Net have the same inference procedure.

V. Experiments

A. Datasets and Metrics

Following previous works, MixATIS and MixSNIPS [10], [37], [38] are taken as testbeds. MixATIS includes 13,162 utterances for training, 756 ones for validation and 828 ones for testing. MixSNIPS includes 39,776 utterances for training, 2,198 ones for validation and 2,199 ones for testing.

As for evaluation metrics, following previous works, we adopt accuracy (Acc) for multiple intent detection, F1 score for slot filling, and overall accuracy (Acc) for the sentence-level semantic frame parsing. Overall accuracy denotes the ratio of sentences whose intents and slots are all correctly predicted.

B. Implementation Details

We construct several group of our models, which are based on LSTM encoder and pre-trained language model (PTLM) encoders (e.g. BERT [41], RoBERTa [42], XLNet [43]).

LSTM: Following previous works, the word and label embeddings are trained from scratch. The dimensions of word embedding, label embedding, and hidden state are 256 on MixATIS, while on MixSNIPS they are 256, 128, and 256. The layer number of all GNNs is 2. Adam [44] is used to train our model with a learning rate of $1e^{-3}$ and a weight decay of $1e^{-6}$. As for the coefficients (14), γ is 0.9 on MixATIS and 0.8 on MixSNIPS; on both datasets, β_I is $1e^{-6}$ and β_S is $1e^{0}$. The above hyper-parameter settings are for both of Co-guiding Net and Co-guiding-SCL Net. The hyper-parameters for the contrastive learning mechanisms in Co-guiding-SCL Net are set as follows. τ is 0.07. η_I and η_S are 0.1 and 0.01.

PTLM: The models based on PTLM encoders replace the self-attentive encoder with the PTLM encoder. We adopt the base version for each PTLM encoder. The learning rate is set to 1e-5 (tuning from [5e-6, 1e-5, 3e-5, 5e-5]) and adopt AdamW optimizer with default configuration. Hidden state dimension is 768. All other hyper-parameter setting are the same as LSTM-based model.

The model performing best on the dev set is selected then we report its results on the test set. All experiments are conducted on RTX 6000 and DGX-A100 server.

C. Baselines

We compare our LSTM-based Co-guiding Net and Co-guiding-SCL Net with Attention BiRNN [39], Slot-Gated [3], SF-ID [6], Stack-Propagation [7], Joint Multiple ID-SF [9], AGIF [10] and GL-GIN [12]. We reproduce the results of GL-GIN using its official source code and default hyper-parameter setting. We compare our PTLM-based Co-guiding Net and Co-guiding-SCL Net with PTLM-based GL-GIN, which is implemented by our self. For fair comparison, we adopt the same learning rate and optimizer setting with our PTLM-based models. And other hyper-parameter setting are the same with GL-GIN.

D. Main Results

The performances of our models and baselines are shown in Table III, from which we have the following observations:

1) Comparison of Our Models and Baselines: (1) Co-guiding Net and Co-guiding-SCL Net gain significant and consistent improvements over baselines on all tasks and datasets. Specifically, on MixATIS dataset, compared with GL-GIN, Co-guiding-SCL Net achieves significant improvements of 21.3%, 2.4%, and 4.1% on sentence-level semantic frame parsing, slot filling, and multiple intent detection, respectively; on MixSNIPS dataset, it overpasses GL-GIN by 5.3%, 1.2% and 1.8% on sentence-level semantic frame parsing, slot filling and multiple intent detection, respectively. The promising results of our model can be attributed to the mutual guidances between multiple intent detection and slot filling, allowing the two tasks to provide crucial clues for each other. Besides, our designed HSLGs and HGATs can effectively model the interactions among the semantics nodes and label nodes, extracting the indicative clues from initial predictions. And our proposed single-task supervised contrastive learning and co-guiding supervised contrastive learning can further capture single-task and dual-task semantics contrastive relations.

(2) Our models achieve larger improvements on multiple intent detection than slot filling. The reason is that except for the guidance from multiple intent detection to slot filling, our models also achieve the guidance from slot filling to multiple intent detection, while previous models all ignore this. Besides, previous methods model the semantics-label interactions by homogeneous graph and GAT, limiting the performance. Differently, our model uses the heterogeneous semantics-label graphs to represent different relations among the semantic nodes and the label nodes, then applies the proposed HGATs over the graphs to achieve the interactions. Consequently, their performances (especially on multiple intent detection) are significantly inferior to our model.

(3) The improvements in overall accuracy are much sharper. We suppose the reason is that the achieved mutual guidances make the two tasks deeply coupled and allow them to stimulate each other using their initial predictions. For each task, its final outputs are guided by its and another task's initial predictions. By this means, the correct predictions of the two tasks can be better aligned. As a result, more test samples get correct sentence-level semantic frame parsing results, and then overall accuracy is boosted.

(4) Based on PTLM encoders, our model brings more significant improvements than GL-GIN. This is because GL-GIN performs semantic interactions, while PTLMs have strong abilities on semantics. Differently, our models make the first attempt to achieve the semantics-label interactions, which cannot be achieved by PTLMs. Therefore, our models' advantages do not overlap with PTLMs', and the high-quality semantics representations generated by PTLM can cooperate well with the co-guiding mechanism of our model.

2) Comparison of Co-Guiding Net and Co-Guiding-SCL Net: The performance improvements of Co-guiding-SCL Net over Co-guiding Net come from our proposed single-task supervised contrastive learning and co-guiding supervised contrastive learning. And we can observe that Co-guiding-SCL Net can obtain larger improvements based on PTLM encoders than the LSTM encoder. We suspect the reason is that PTLM can generate much higher-quality semantics representations than LSTM, and then the contrastive learning mechanisms can be improved because they are performed on the representations.

E. Model Analysis of Co-Guiding Net

We conduct a set of ablation experiments to verify the advantages of our work from different perspectives, and the results are shown in Table IV.

1) Effect of Slot-to-Intent Guidance: One of the core contributions of our work is achieving the mutual guidances between

TABLE III
RESULTS COMPARISON

LSTM-based Models	MixATIS			MixSNIPS		
	Overall(Acc)	Slot (F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)
Attention BiRNN [39]	39.1	86.4	74.6	59.5	89.4	95.4
Slot-Gated [3]	35.5	87.7	63.9	55.4	87.9	94.6
Bi-Model [40]	34.4	83.9	70.3	63.4	90.7	95.6
SF-ID [6]	34.9	87.4	66.2	59.9	90.6	95.0
Stack-Propagation [7]	40.1	87.8	72.1	72.9	94.2	96.0
Joint Multiple ID-SF [9]	36.1	84.6	73.4	62.9	90.6	95.1
AGIF [10]	40.8	86.7	74.4	74.2	94.2	95.1
GL-GIN [12]	42.8 (±0.20)	87.9 (±0.36)	76.0 (±0.36)	73.2 (±0.42)	93.9 (±0.12)	95.8 (±0.31)
Co-guiding Net (ours)	**50.9**† (±0.47)	**89.5**† (±0.49)	**78.7**† (±0.32)	**77.2**† (±0.41)	**94.9**† (±0.20)	**97.5**† (±0.16)
Co-guiding-SCL Net (ours)	**51.9**† (±0.18)	**90.0**† (±0.14)	**79.1**† (±0.18)	**77.1**† (±0.43)	**95.0**† (±0.10)	**97.5**† (±0.14)
BERT-based Models	**MixATIS**			**MixSNIPS**		
	Overall(Acc)	Slot (F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)
BERT+Linear	47.9 (±0.20)	87.0 (±0.35)	80.6 (±0.67)	84.6(±0.48)	96.8 (±0.23)	97.4 (±0.30)
BERT+GL-GIN [12]	49.3 (±0.82)	86.7 (±0.51)	79.5 (±0.21)	84.9 (±0.40)	96.8 (±0.18)	96.9 (±0.23)
BERT+Co-guiding Net (ours)	**52.4**† (±0.32)	**88.3**† (±0.43)	**82.3**† (±0.32)	**86.4**† (±0.21)	**97.1**† (±0.18)	**97.4**† (±0.13)
BERT+Co-guiding-SCL Net (ours)	**54.0**† (±0.44)	**89.1**† (±0.17)	**84.2**† (±0.18)	**87.4**† (±0.12)	**97.3**† (±0.02)	**98.2**† (±0.09)
RoBERTa-based Models	**MixATIS**			**MixSNIPS**		
	Overall(Acc)	Slot (F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)
RoBERTa+Linear	48.4(±0.32)	86.0 (±0.32)	80.3 (±0.49)	82.1(±0.32)	96.0 (±0.24)	97.4 (±0.07)
RoBERTa+GL-GIN [12]	49.9 (±0.35)	86.8 (±0.37)	80.8(±0.19)	82.5 (±0.36)	96.3 (±0.64)	97.3 (±0.40)
RoBERTa+Co-guiding Net (ours)	**54.3**† (±0.41)	**88.4**† (±0.35)	**83.2**† (±0.30)	**83.9**† (±0.30)	**97.5**† (±0.17)	**98.0**† (±0.20)
RoBERTa+Co-guiding-SCL Net (ours)	**56.6**† (±0.43)	**89.6**† (±0.16)	**84.4**† (±0.32)	**85.2**† (±0.07)	**98.5**† (±0.06)	**98.3**† (±0.07)
XLNet-based Models	**MixATIS**			**MixSNIPS**		
	Overall(Acc)	Slot (F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)
XLNet+Linear	52.5(±0.24)	87.5 (±0.30)	82.2 (±0.43)	84.3(±0.25)	96.6 (±0.20)	97.2 (±0.09)
XLNet+GL-GIN [12]	53.1(±0.13)	87.8 (±0.24)	82.6 (±0.27)	84.8(±0.28)	96.6 (±0.27)	96.9 (±0.09)
XLNet+Co-guiding Net (ours)	**54.0**† (±0.20)	**88.6**† (±0.15)	**83.8**† (±0.30)	**86.1**† (±0.18)	**97.1**† (±0.13)	**97.6**† (±0.11)
XLNet+Co-guiding-SCL Net (ours)	**56.7**† (±0.72)	**89.2**† (±0.21)	**84.6**† (±0.28)	**87.7**† (±0.19)	**97.6**† (±0.09)	**98.7**† (±0.09)

We report the average results of three runs with different random seeds. ± denotes standard deviation. † denotes our model significantly outperforms baselines with $p < 0.01$ under t-test.
Best scores are in bold.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS

Models	MixATIS			MixSNIPS		
	Overall(Acc)	Slot (F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)
Co-guiding Net	**50.9**	**89.5**	**78.7**	**77.2**	**94.9**	**97.5**
w/o S2I-guidance	47.4 (↓3.5)	88.5 (↓1.0)	76.8 (↓1.9)	76.3 (↓0.9)	94.5 (↓0.4)	96.7 (↓0.8)
w/o I2S-guidance	47.3 (↓3.6)	88.4 (↓1.1)	77.2 (↓1.5)	76.2 (↓1.0)	94.7 (↓0.2)	97.3 (↓0.2)
w/o relations	45.6 (↓5.3)	88.0 (↓1.5)	77.5 (↓1.2)	76.0 (↓1.2)	94.5 (↓0.4)	97.0 (↓0.5)
+ Local Slot-aware GAT	50.7 (↓0.2)	89.2 (↓0.3)	78.6 (↓0.1)	75.6 (↓1.6)	94.5 (↓0.4)	96.1 (↓1.4)

Best scores are in bold.

multiple intent detection and slot filling, while previous works only leverage the one-way message from intent to slot. Therefore, compared with previous works, one of the advantages of our work is modeling the slot-to-intent guidance. To verify this, we design a variant termed *w/o S2I-guidance* and its result is shown in Table IV. We can observe that Intent Acc drops by 2.0% on MixATIS and 0.8% on MixSNIPS. Moreover, Overall Acc drops more significantly: 3.6% on MixATIS and 0.9% on MixSNIPS. This proves that the guidance from slot to intent can effectively benefit multiple intent detection, and achieving the mutual guidances between the two tasks can significantly improve Overall Acc.

Besides, although both of *w/o S2I-guidance* and GL-GIN only leverage the one-way message from intent to slot, *w/o S2I-guidance* outperforms GL-GIN by large margins. We attribute this to our proposed heterogeneous semantics-label graphs and heterogeneous graph attention networks, whose advantages are verified in Section V-E3.

2) Effect of Intent-to-Slot Guidance: To verify the effectiveness of intent-to-slot guidance, we design a variant termed *w/o I2S-guidance* and its result is shown in Table IV. We can find that the intent-to-slot guidance has a significant impact on performance. Specifically, *w/o I2S-guidance* cause nearly the same extent of performance drop on Overall Acc, proving that both of the intent-to-slot guidance and slot-to-intent guidance are indispensable and achieving the mutual guidances can significantly boost the performance.

3) Effect of HSLGs and HGATs: In this paper, we design two HSLGs: (i.e., S2I-SLG, I2S-SLG) and two HGATs (i.e., S2I-HGAT, I2S-HGAT). To verify their effectiveness, we design a variant termed *w/o relations* by removing the relations on the two HSLGs. In this case, S2I-SLG/I2S-SLG collapses to a homogeneous graph, and S2I-HGAT/I2S-HGAT collapses to a general GAT based on multi-head attentions. From Table IV, we can observe that *w/o relations* obtains dramatic drops on all metrics on both datasets. The apparent performance gap between *w/o relations* and Co-guiding Net verifies that (1) our proposed HSLGs can effectively represent the different relations among the semantics nodes and label nodes, providing appropriate platforms for modeling the mutual guidances between the two tasks; (2) our proposed HGATs can sufficiently and effectively model interactions between the semantics and indicative label information via achieving the relation-specific attentive information aggregation on the HSLGs.

Besides, although *w/o relations* obviously underperforms Co-guiding Net, it still significantly outperforms all baselines. We attribute this to the fact that our model achieves the mutual guidances between the two tasks, which allows them to promote each other via cross-task correlations.

4) Effect of I2S-HGAT for Capturing Local Slot Dependencies: [12] propose a Local Slot-aware GAT module to alleviate the uncoordinated slot problem (e.g., *B-singer* followed by *I-song*) [17] caused by the non-autoregressive fashion of slot filling. And the ablation study in [12] proves that this module effectively improves the slot filling performance by modeling the local dependencies among slot hidden states. In their model (GL-GIN), the local dependencies are modeled in both of the local slot-aware GAT and subsequent global intent-slot GAT. We suppose the reason why GL-GIN needs the local Slot-aware GAT is that the global intent-slot GAT in GL-GIN cannot effectively capture the local slot dependencies. GL-GIN's global slot-intent graph is homogeneous, and the GAT working on it treats the slot semantics nods and the intent label nodes equally without discrimination. Therefore, each slot hidden state receives indiscriminate information from both of its local slot hidden states and all intent labels, making it confusing to capture the local slot dependencies. In contrast, we believe our I2S-HLG and I2S-HGAT can effectively capture the slot local dependencies along the specific *slot_semantics_dependencies* relation, which is modeled together with other relations. Therefore, our Co-guiding Net does not include another module to capture the slot local dependencies.

To verify this, we design a variant termed *+Local Slot-aware GAT*, which is implemented by augmenting Co-guiding Net with the Local Slot-aware GAT [12] located after the Intent-aware BiLSTM$_s$ (the same position with GL-GIN). And its result is shown in Table IV. We can observe that not only the Local Slot-aware GAT does not bring improvement, it even causes performance drops. This proves that our I2S-HGAT can effectively capture the local slot dependencies.

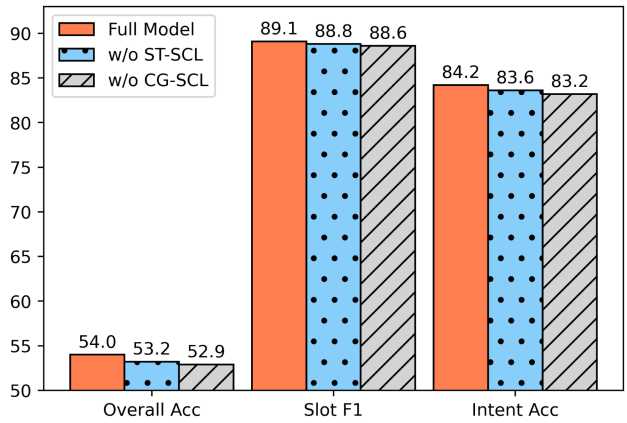


Fig. 8. Ablation results. Full model denotes BERT+Co-guiding-SCL Net. w/o ST-SCL denotes the single-task supervised contrastive learning is removed. w/o CG-SCL denotes the co-guiding supervised contrastive learning is removed.

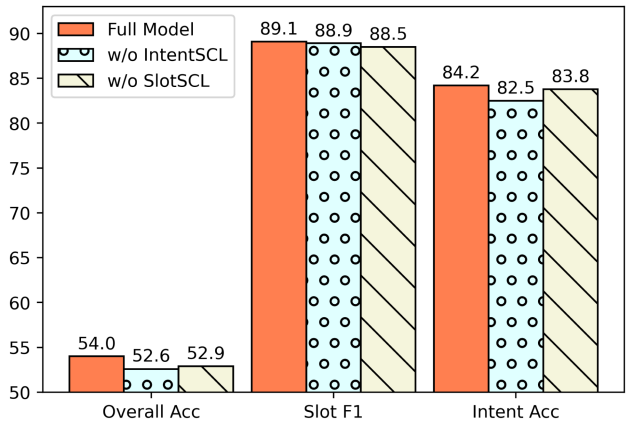


Fig. 9. Ablation results. Full model denotes BERT+Co-guiding-SCL Net. w/o IntentSCL denotes the single-task and S=slot-guided multi-intent supervised contrastive learning mechanism are removed. w/o SlotSCL denotes the single-task and intent-guided slot supervised contrastive learning mechanisms are removed.

F. Analysis of Supervised Contrastive Learning Mechanisms in Co-Guiding-SCL Net

We conduct a set of ablation to verify the advantages of our proposed supervised contrastive learning mechanisms in Co-guiding-SCL Net, and the results on MixATIS dataset are shown in Figs. 8 and 9.

1) Effect of Single-Task/Co-Guiding Supervised Contrastive Learning: From Fig. 8 we can observe that removing single-task supervised contrastive learning (ST-SCL) leads to performance decrease. This is because ST-SCL can improve the label distribution of the initial estimations in the first stage via leveraging the single-task semantics contrastive relations, which is achieved

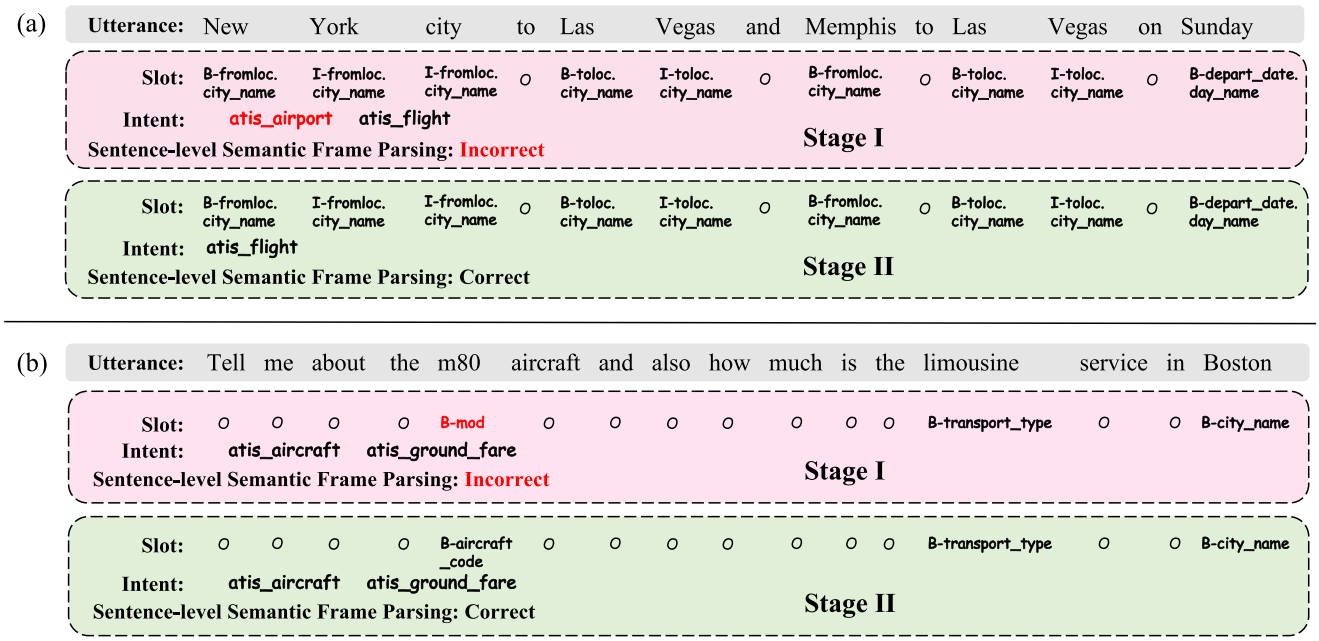


Fig. 10. Case study of slot-to-intent guidance (a) and intent-to-slot guidance (b). Red color denotes error.

by drawing together the representations corresponding to the same/similar labels while pushing apart the ones corresponding to different labels. And the better label distributions can provide more reliable indicative information for the dual-task co-guiding mechanism in the second stage.

We can find that the variant without co-guiding supervised contrastive learning (CG-SCL) perform significantly worse than the full model. This proves the advantages of CG-SCL, which can further capture dual-task semantics contrastive relations at the second stage via integrating both tasks' labels as supervision signals for the supervised contrastive learning mechanism. In the second stage, CG-SCL cooperates with the HGATs to comprehensively and effectively model the dual-task mutual guidances, significantly improving the final predictions.

2) Effect of Intent/Slot Supervised Contrastive Learning: From Fig. 9 we can observe that removing intent supervised contrastive learning (IntentSCL) leads to the performance decreases on intent accuracy, while causes the model performs worse on slot filling and sentence-level semantics parsing at the same time. And removing slot supervised contrastive learning (SlotSCL) leads to the performance decreases not only on slot F1, but also on intent accuracy and overall accuracy. There are two reasons. First, IntentSCL and SlotSCL can effectively improve the model performances on multiple intent detection and slot filling, respectively. Second, the co-guiding supervised contrastive learning further makes the two tasks deeply coupled and interrelated on each other's performances. Therefore, removing anyone of IntentSCL and SlotSCL leads to the decreases on all of overall accuracy, slot F1 and intent accuracy.

G. Case Study

To demonstrate how our model allows the two tasks to guide each other, we present two cases in Fig. 10.

TABLE V
COMPARISON WITH SOTA ON TRAINING TIME AND LATENCY

Models	Training Time per Epoch	Latency /Inference Time per Utterance
GL-GIN	68s	2.6ms
Co-guiding Net	70s	2.9ms
Co-guiding-SCL Net	82s	2.9ms
BERT+GL-GIN	148s	5.6ms
BERT+Co-guiding Net	156s	6.0ms
BERT+Co-guiding-SCL Net	185s	6.0ms

1) Slot-to-Intent Guidance: From Fig. 10(a), we can observe that in the first stage, all slots are correctly predicted, while multiple intent detection obtains a redundant intent `atis_airport`. In the second stage, our proposed S2I-HGAT operates on S2I-HLG. It aggregates and analyzes the slot label information from the slot predictions of the first stage, extracting the indicative information that most slot labels are about `city_name` while no information about *airport* is mentioned. Then this beneficial guidance information is passed into intent semantics nodes whose representations are then fed to the intent decoder for prediction. In this way, the guidance from slot filling helps multiple intent detection predict correctly.

2) Intent-to-Slot Guidance: In the example shown in Fig. 10(b), in the first stage, correct intents are predicted, while there is an error in the predicted slots. In the second stage, our proposed I2S-HGAT operates on I2S-HLG. It comprehensively analyzes the indicative information of *airecraft* from both of slot semantics node *aircraft* and intent label node `atis_aircraft`. Then this beneficial guidance information is passed into the slot semantics of *m80*, whose slot is therefore correctly inferred.

TABLE VI
PERFORMANCES BASED ON MBERT

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
GL-CLEF	70.77	70.74	69.86	68.77	72.23	70.65	70.85	69.93	70.77	70.51
Co-guiding Net (ours)	**97.65**	95.70	96.02	94.10	71.00	65.43	94.10	67.37	76.33	84.19
Co-guiding-SCL Net (ours)	97.54	96.52	96.56	94.69	75.92	64.56	94.21	72.07	76.19	85.36
GL-CLEF+Co-guiding Net (ours)	97.54	**96.82**	**97.25**	96.80	76.22	73.33	95.29	78.32	82.98	88.28
GL-CLEF+Co-guiding-SCL Net (ours)	97.54	96.52	96.72	**97.13**	**80.74**	**75.73**	**95.40**	**78.79**	**83.54**	**89.12**
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
GL-CLEF	94.85	85.77	85.51	84.99	56.52	65.92	81.08	65.66	78.47	77.64
Co-guiding Net (ours)	96.28	81.57	83.63	81.38	41.06	39.06	74.00	51.93	64.21	68.12
Co-guiding-SCL Net (ours)	**96.30**	79.76	83.19	81.77	31.98	28.37	75.09	58.23	64.58	66.59
GL-CLEF+Co-guiding Net (ours)	95.84	84.64	85.34	84.80	58.62	**66.05**	**81.35**	66.84	**79.78**	78.14
GL-CLEF+Co-guiding-SCL Net (ours)	95.76	**86.92**	**86.12**	**85.86**	**59.87**	64.55	81.15	**68.92**	79.14	**78.70**
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
GL-CLEF	66.55	48.69	46.37	45.98	18.92	**33.52**	45.85	23.26	42.40	41.28
Co-guiding Net (ours)	**89.06**	56.88	57.15	52.89	9.63	8.80	46.11	15.57	28.59	40.52
Co-guiding-SCL Net (ours)	88.88	55.12	56.83	54.15	6.57	7.37	47.12	17.62	27.36	40.11
GL-CLEF+Co-guiding Net (ours)	88.24	64.87	62.07	61.27	20.34	27.58	59.30	29.09	**51.81**	51.62
GL-CLEF+Co-guiding-SCL Net (ours)	87.76	**67.90**	**63.30**	**62.16**	**22.62**	30.66	**59.94**	**31.89**	**51.81**	**53.11**

* denotes our model significantly outperforms baselines with $p < 0.05$ under the t-test. Best scores are in bold.

TABLE VII
PERFORMANCES BASED ON XLM-R

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
GL-CLEF	70.77	70.74	69.86	68.77	72.23	70.65	70.85	69.93	70.77	70.51
Co-guiding Net (ours)	97.72	94.17	97.09	96.29	86.41	73.55	**96.79**	66.95	78.61	87.51
Co-guiding-SCL Net (ours)	97.80	94.58	97.29	95.20	84.06	68.92	96.41	55.99	83.91	86.02
GL-CLEF+Co-guiding Net (ours)	97.50	95.93	97.25	96.33	**87.64**	73.85	95.55	**75.76**	87.42	**89.69**
GL-CLEF+Co-guiding-SCL Net (ours)	**98.17**	**96.15**	**97.83**	**97.01**	87.23	**77.35**	96.04	74.87	89.59	90.47
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
GL-CLEF	96.01	85.30	86.85	83.00	63.95	67.92	79.99	56.24	**80.82**	77.78
Co-guiding Net (ours)	96.08	82.51	85.34	80.16	58.61	29.83	**80.33**	37.99	61.98	68.09
Co-guiding-SCL Net (ours)	**96.20**	83.95	85.50	81.60	63.70	32.68	80.04	42.29	57.31	69.25
GL-CLEF+Co-guiding Net (ours)	95.84	85.49	**87.29**	81.71	**72.05**	**72.22**	79.08	56.10	79.11	**78.76**
GL-CLEF+Co-guiding-SCL Net (ours)	95.90	**86.59**	87.03	**81.84**	71.01	69.49	79.79	**56.79**	80.11	78.73
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
GL-CLEF	66.48	48.28	46.78	45.85	22.28	28.48	42.30	15.94	44.53	40.10
Co-guiding Net (ours)	88.43	60.54	61.50	54.07	23.33	5.94	**57.66**	5.13	22.36	42.11
Co-guiding-SCL Net (ours)	**88.69**	60.84	61.13	56.55	28.82	7.04	57.03	5.64	18.55	42.70
GL-CLEF+Co-guiding Net (ours)	87.98	64.24	63.59	58.20	**38.75**	**35.52**	55.87	19.81	53.83	53.09
GL-CLEF+Co-guiding-SCL Net (ours)	88.50	**66.59**	**63.76**	**58.87**	35.83	34.73	56.99	**20.14**	**56.51**	**53.55**

* denotes our model significantly outperforms baselines with $p < 0.05$ under the t-test. Best scores are in bold.

H. Computation Efficiency

The training time and latency of our models and the state-of-the-art model are shown in Fig. 5. We can find that our Co-guiding-SCL Net costs some more training time due to the contrastive learning operations. As for latency, both our Co-guiding Net and Co-guiding-SCL Net are comparable to GL-GIN, while they can significantly outperform it. Our proposed contrastive learning mechanisms in Co-guiding-SCL only work in the training process, without affecting the latency.

I. Zero-Shot Cross-Lingual Multi-Intent SLU

1) Experiment Setup: Dataset and Metrics We evaluate our model on the multilingual benchmark dataset of MultiATIS++ [45]. This dataset includes the multi-intent training samples in English and the testing samples in 9 languages: English (en), Spanish (es), Portuguese (pt), German (de), French (fr), Chinese (zh), Japanese (ja), Hindi (hi) and Turkish (tr). And intent accuracy (Acc), slot filling and overall accuracy (Acc) are adopted as the evaluation metrics.

Baseline and Implementation: Currently, the state-of-the-art model for zero-shot cross-lingual SLU is GL-CLEF [35], which utilizes the unsupervised contrastive learning to align the source language semantics and the target language semantics. However, it is designed for single-intent SLU. Therefore, we modify its official code to make it available for multi-intent SLU. We use the sigmoid function and a linear layer, which is similar to (4), to replace its original intent classification module. And we replace

its original loss function with the loss function (11) used in our models.

Apart from evaluating the performances of Co-guiding Net and Co-guiding-SCL Net, we also combine them with GL-CLEF, forming GL-CLEF+Co-guiding Net and Co-guiding-SCL Net. For fair comparison, the hyper-parameters of GL-CLEF used in GL-CLEF, GL-CLEF+Co-guiding Net and Co-guiding-SCL Net is directly retrieved from its original paper and the official code. As for the hyper-parameters of Co-guiding Net and Co-guiding-SCL Net, we just use the ones staged in Section V-B.

We conduct two groups of experiments, which are based on two multilingual pre-trained language models (e.g., mBERT [41] and XLM-R [46]), respectively. And we report the average results of three runs with different random seeds.

2) Results Analysis: The results of the models based on mBERT and XLM-R are shown in Tables VI and VII. First, we can observe that although Co-guiding Net and Co-guiding-SCL Net obtain significant improvements on English, it is hard to say that Co-guiding Net and Co-guiding-SCL Net can outperform GL-CLEF on the average overall accuracy of the total 9 languages. We suspect the reason is that our models have a strong ability to model the mutual guidances between the two tasks and the semantics-label interactions, while there is no multilingual module in our models, which makes it hard to transfer the learned beneficial knowledge from the source language (English) to target languages. However, if combining our models with GL-CLEF, we can observe obvious improvements. Specifically, based on XLM-R, GL-CLEF+Co-guiding-SCL Net gains a relative improvement of 33.5% over GL-CLEF on the average overall accuracy of the total 9 languages. There are two reasons. First, our models' advantage is capturing the beneficial knowledge learned from the source language via modeling the dual-task mutual guidances and capturing the fine-grained dual-task correlations. Second, the semantics alignments of GL-CLEF work as a bridge that can transfer the knowledge learned from the source language to the target language. Therefore, our models work well with GL-CLEF, achieving promising results for zero-shot cross-lingual multi-intent SLU.

VI. Conclusion

In this paper, we propose a novel two-stage framework that allows the two tasks to guide each other in the second stage using the predicted labels at the first stage. Based on this framework, we propose two novel models: Co-guiding Net and Co-guiding-SCL Net. To represent the relations among the semantics node and label nodes, we propose two heterogeneous semantics-label graphs and two heterogeneous graph attention networks to model the mutual guidances between intents and slots. Besides, we propose the single-task supervised contrastive learning and co-guiding supervised contrastive learning, which are performed in the first stage and second stage, receptively. We conduct extensive experiments to evaluate our models on multi-intent SLU and zero-shot cross-lingual multi-intent SLU. Experiment results on benchmark datasets show that our model significantly outperforms previous models by large margins. On multi-intent SLU task, our model obtains a relative improvement of 21.3% over the previous best model on MixATIS dataset. On zero-shot cross-lingual multi-intent SLU task, our model can relatively improve the state-of-the-art model by 33.5%on average in terms of overall accuracy for the total 9 languages.

In addition, this work provides some general insights of exploiting the word-level and sentence-level semantics correlations via leveraging the dual-task contrastive relations among the word-level and sentence-level labels. This idea can be applied to other scenarios which jointly tackle the sentence-level and word-level tasks.

References

[1] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "POMDP-based statistical spoken dialog systems: A review," *Proc. IEEE*, vol. 101, no. 5, pp. 1160–1179, May 2013.

[2] G. Tur and R. De Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information From Speech*. Hoboken, NJ, USA: Wiley, 2011.

[3] C.-W. Goo et al., "Slot-gated modeling for joint slot filling and intent prediction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, New Orleans, Louisiana, Association for Computational Linguistics, 2018, pp. 753–757.

[4] C. Li, L. Li, and J. Qi, "A self-attentive model with gate mechanism for spoken language understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Association for Computational Linguistics, 2018, pp. 3824–3833.

[5] Y. Liu, F. Meng, J. Zhang, J. Zhou, Y. Chen, and J. Xu, "CM-Net: A novel collaborative memory network for spoken language understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, Association for Computational Linguistics, 2019, pp. 1051–1060.

[6] H. E., P. Niu, Z. Chen, and M. Song, "A novel bi-directional interrelated model for joint intent detection and slot filling," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Association for Computational Linguistics, 2019, pp. 5467–5471.

[7] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, Association for Computational Linguistics, 2019, pp. 2078–2087.

[8] B. Kim, S. Ryu, and G. G. Lee, "Two-stage multi-intent detection for spoken language understanding," *Multimedia Tools Appl.*, vol. 76, no. 9, pp. 11377–11390, 2017.

[9] R. Gangadharaiah and B. Narayanaswamy, "Joint multiple intent detection and slot labeling for goal-oriented dialog," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 564–569.

[10] L. Qin, X. Xu, W. Che, and T. Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 1807–1816.

[11] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018.

[12] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, "GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, Association for Computational Linguistics, 2021, pp. 178–188.

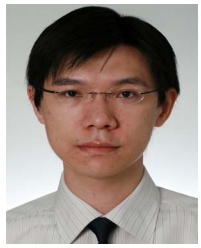
[13] B. Xing and I. Tsang, "Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, 2022, pp. 159–169.

[14] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, New York, NY, USA, 2016, pp. 2993–2999.

[15] D. Hakkani-Tür et al., "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 715–719.

[16] C. Zhang, Y. Li, N. Du, W. Fan, and P. Yu, "Joint slot filling and intent detection via capsule neural networks," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Association for Computational Linguistics, 2019, pp. 5259–5267.
[17] D. Wu, L. Ding, F. Lu, and J. Xie, "SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Association for Computational Linguistics, 2020, pp. 1932–1937.
[18] L. Qin, T. Liu, W. Che, B. Kang, S. Zhao, and T. Liu, "A co-interactive transformer for joint slot filling and intent detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 8193–8197.
[19] J. Ni, T. Young, V. Pandelea, F. Xue, V. Adiga, and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey," 2021, *arXiv:2105.04387*.
[20] Y. Cao, Z. Liu, C. Li, Z. Liu, J. Li, and T.-S. Chua, "Multi-channel graph neural network for entity alignment," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Association for Computational Linguistics, 2019, pp. 1452–1461.
[21] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2020, pp. 3229–3238.
[22] J. Shi, S. Cao, L. Hou, J. Li, and H. Zhang, "TransferNet: An effective and transparent framework for multi-hop question answering over relation graph," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Punta Cana, Dominican Republic:Association for Computational Linguistics, 2021, pp. 4149–4158.
[23] B. Xing and I. W. Tsang, "Understand me, if you refer to aspect knowledge: Knowledge-aware gated recurrent memory network," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1092–1102, Oct. 2022.
[24] B. Xing and I. Tsang, "DARER: Dual-task temporal relational recurrent reasoning network for joint dialog sentiment classification and act recognition," in *Proc. Findings Assoc. Comput. Linguistics*. Dublin, Ireland, Association for Computational Linguistics, May 2022, pp. 3611–3621.
[25] B. Xing and I. Tsang, "DigNet: Digging clues from local-global interactive graph for aspect-level sentiment classification," 2022, *arXiv:2201.00989*.
[26] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 4568–4578.
[27] B. Xing and I. Tsang, "Neural subgraph explorer: Reducing noisy information via target-oriented syntax graph pruning," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 4425–4431.
[28] B. Xing and I. W. Tsang, "Co-evolving graph reasoning network for emotion-cause pair extraction," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases: Res. Track*, Cham, Switzerland, Springer Nature, 2023, pp. 305–322.
[29] B. Xing and I. W. Tsang, "Relational temporal graph reasoning for dual-task dialogue language understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13170–13184, Nov. 2023.
[30] M. S. Schlichtkrull et al., "Modeling relational data with graph convolutional networks," in *Proc. Semantic Web 15th Int. Conf.*, 2018, pp. 593–607.
[31] B. Xing and I. Tsang, "Group is better than individual: Exploiting label topologies and label relations for joint multiple intent detection and slot filling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, 2022, pp. 3964–3975. [Online]. Available: https://aclanthology.org/2022.emnlp-main.263
[32] D. Zhang et al., "Pairwise supervised contrastive learning of sentence representations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, pp. 5786–5798.
[33] Y. Zhou, P. Liu, and X. Qiu, "KNN-contrastive learning for out-of-domain intent classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, Association for Computational Linguistics, 2022, pp. 5129–5141.
[34] Z. Wang, P. Wang, L. Huang, X. Sun, and H. Wang, "Incorporating hierarchy into text encoder: A contrastive learning approach for hierarchical text classification," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 7109–7119.
[35] L. Qin et al., "GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, Association for Computational Linguistics, 2022, pp. 2677–2686.
[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[37] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Proc. Speech Natural Lang.: Proc. Workshop Held Hidden Valley*, Pennsylvania, 1990.
[38] A. Coucke et al., "Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces," 2018, *ArXiv:1805.10190*.
[39] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 685–689.
[40] Y. Wang, Y. Shen, and H. Jin, "A bi-model based RNN semantic frame parsing model for intent detection and slot filling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, New Orleans, Louisiana, Association for Computational Linguistics, 2018, pp. 309–314.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[42] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv: 1907.11692*.
[43] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015.
[45] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Association for Computational Linguistics, 2020, pp. 5052–5063.
[46] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2020, pp. 8440–8451.

Bowen Xing received the BE and master's degrees from the Beijing Institute of Technology, Beijing, China, in 2017 and 2020, respectively. He is currently working toward the PhD degree with the Australian Artificial Intelligence Institute (AAII), University of Technology Sydney (UTS). His research focuses on graph neural networks, multi-task learning, sentiment analysis, and dialog system.



Ivor W. Tsang (Fellow, IEEE) is the director of A*STAR Centre for Frontier AI Research (CFAR). Previously, he was a professor of artificial intelligence with the University of Technology Sydney (UTS), and a research director of the Australian Artificial Intelligence Institute (AAII). His research focuses on transfer learning, deep generative models, learning with weakly supervision, Big Data analytics for data with extremely high dimensions in features, samples and labels. His work is recognised internationally for its outstanding contributions to those fields. In 2013, he received his ARC Future Fellowship for his outstanding research on Big Data analytics and large-scale machine learning. In 2019, his JMLR paper "Towards ultrahigh dimensional feature selection for Big Data" received the International Consortium of Chinese Mathematicians Best Paper Award. In 2020, he was recognized as the AI 2000 AAAI/IJCAI Most Influential Scholar in Australia for his outstanding contributions to the field, between 2009 and 2019. His research on transfer learning was awarded the Best Student Paper Award at CVPR 2010 and the 2014 IEEE TMM Prize Paper Award. In addition, he received the IEEE TNN Outstanding 2004 Paper Award in 2007 for his innovative work on solving the inverse problem of non-linear representations. Recently, he was conferred the IEEE fellow for his outstanding contributions to large-scale machine learning and transfer learning. He serves as the editorial board for the *Journal of Machine Learning Research*, *The Modern Language Journal*, *Journal of Artificial Intelligence Research*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Artificial Intelligence*, *IEEE Transactions on Big Data*, and *IEEE Transactions on Emerging Topics in Computational Intelligence*. He serves as a senior area chair/area chair for NeurIPS, ICML, AAAI and IJCAI, and the steering committee of ACML.