PCAD: Towards ASR-Robust Spoken Language Understanding via Prototype Calibration and Asymmetric Decoupling

Xianwei Zhuang, Xuxin Cheng, Liming Liang, Yuxin Xie,
Zhichang Wang, Zhiqi Huang, Yuexian Zou*
ADSPLAB, School of ECE, Peking University, China
{xwzhuang,chengxx,limingliang,yuxinxie,wzcc}@stu.pku.edu.cn,
{zhiqihuang, zouyx}@pku.edu.cn

Abstract

Spoken language understanding (SLU) inevitably suffers from error propagation from automatic speech recognition (ASR) in actual scenarios. Some recent works attempt to alleviate this issue through contrastive learning. However, they (1) sample negative pairs incorrectly in pre-training; (2) only focus on implicit metric learning while neglecting explicit erroneous predictions; (3) treat manual and ASR transcripts indiscriminately. In this paper, we propose a novel framework termed **PCAD**, which can calibrate bias and errors and achieve adaptive-balanced decoupling training. Specifically, PCAD utilizes a prototype-based loss to aggregate label and prediction priors and calibrate bias and error-prone semantics for better inter-class discrimination and intra-class consistency. We theoretically analyze the effect of this loss on robustness enhancement. Further, we leverage a teacher-student model for asymmetric decoupling training between different transcripts and formulate a novel gradient-sensitive exponential moving averaging (GS-EMA) algorithm for adaptive balance of accuracy and robustness. Experiments on three datasets show that PCAD significantly outperforms existing approaches and achieves new state-of-the-art performance.

1 Introduction

Spoken Language Understanding (SLU) is the core component of intelligent assistants, aiming to capture the semantics information of user queries (Tur and Mori, 2011). The mainstream SLU solutions are divided into two categories: pipeline and end-to-end methods (Radfar et al., 2020). Pipeline SLU methods combine automatic speech recognition (ASR) and natural language understanding (NLU) in a cascaded manner, enabling them to easily employ pre-trained language models (Cheng et al., 2023a, 2024; Zhuang et al., 2024; Zhu et al., 2024a). However, pipeline SLU suffers from error propagation from ASR in actual scenarios, which inevitably damages the performance of cascaded NLU components.

Some studies explore to address this issue through directly correcting ASR errors (Mani et al., 2020; Wang et al., 2020) or adapting networks via masked language modeling (MLM) (Sundararaman et al., 2021; Huang and Chen, 2019; Zhu et al., 2021), which usually introduces extra speech-related information for correction. Recent works (Chang and Chen, 2022; Cheng et al., 2023b,a) enhance representation through contrastive learning (Chen et al., 2020) to mitigate the impact of ASR errors. Following Chang and Chen (2022); Cheng et al. (2023b,a), we focus on pipeline SLU methods and attempt to improve ASR robustness only by using textual information.

Although these works (Chang and Chen, 2022; Cheng et al., 2023b,a) have achieved promising progress through contrastive learning on ASR and clean transcripts, we discover that they still suffer from (1) **sample bias**: In the label-agnostic pre-training stage, contrastive learning (Chang and Chen, 2022) may mistakenly sample negative pairs (Zhou et al., 2022), resulting in the separation of features with the same semantics (Mishchuk et al., 2017; Zhou et al., 2022). (2) **semantic error**: Chang and Chen (2022); Cheng et al. (2023b,a) attempt to alleviate ASR errors by implementing feature alignment in an implicit space. However, they ignore explicit erroneous intentions and treat correct and incorrect predictions equally during metric learning. (3) **coupling tendency**: Clean and ASR transcripts make different contributions to the accuracy and robustness. Chang and Chen (2022) treat different transcripts indiscriminately and use a mixed set of all transcripts for fine-tuning simultaneously. Despite being aware of this issue, Cheng et al. (2023a) still symmetrically trains different transcripts through dual models.

*Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5235–5246

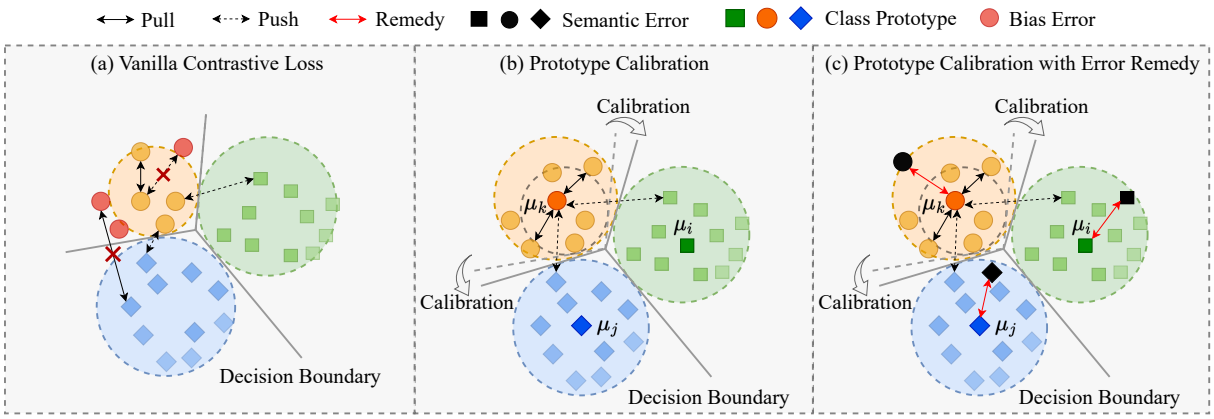


Figure 1: The illustration of the introduced prototype calibration. Figure 1. (a) shows the sampling bias problem encountered in vanilla contrastive learning. Figure 1. (b) shows the prototype-based calibration scheme used in the pre-training phase. Figure 1. (c) shows the prototype-based calibration with semantic error remedy used in the fine-tuning stage.

In this paper, we propose **P**rototype **C**alibration and **A**symmetric **D**ecoupling (**PCAD**) to tackle the above issues. For sample bias, we adopt a prototype contrastive loss as a regularizer to calibrate errors between all features and corresponding prototype centroids (Ye et al., 2019; Li et al., 2020) for alleviating representation deterioration. Class prototypes can be used to include label supervision, thereby alleviating sampling bias caused by label-agnostic pre-training. For the second issue, we introduce an error-sensitive prototype loss which adds a calibration term to the regular prototype loss to punish the misclassified predictions during fine-tuning. We emphasize on learning more accurate prototypes by removing misclassified intentions. Figure 1 illustrates our main idea. We further analyze the theoretical effectiveness of our prototype-based approach for calibrating bias and improving robustness from the perspectives of gradient and robust decision-making in section 3.6. For coupling tendency, we innovatively propose an asymmetric algorithm termed GS-EMA for decoupling training, which utilizes the average gradient saliency to achieve the adaptive balance of accuracy and robustness. To our best knowledge, it is the first to calibrate error-prone SLU representation from a prototype perspective and consider adaptive-balanced decoupling training for improving ASR robustness.

Our contributions can be summarized as: (1) We propose an ASR-robust SLU framework termed PCAD, which introduces a prototype-based calibration loss and its error-sensitive variant for calibrating bias in pre-training and reducing errors in fine-tuning. (2) We introduce a teacher-student model for asymmetric decoupling training and propose a novel GS-EMA algorithm to achieve adaptive balance. (3) Theoretical analysis and extensive experiments on three datasets SLURP (Bastianelli et al., 2020), ATIS (Hemphill et al., 1990) and TREC6 (Li and Roth, 2002) show that PCAD can effectively handle ASR noise and improve robustness in SLU.

2 Related Work

ASR-Robust Spoken Language Understanding. The error propagation of ASR can affect the performance of a series of SLU subtasks. Improving the robustness of SLU tasks to ASR errors has significant practical value. Chang and Chen (2022) first explores utilizing contrastive learning to learn the invariance characteristics between noisy and clean transcripts for improving ASR robustness. Cheng et al. (2023a) realizes the semantic similarity encountered in contrastive learning during the pre-training stage and attempts to alleviate this problem by large-margin contrastive learning (Chen et al., 2021). Cheng et al. (2023b) proposes cross-attention mechanisms to differentiate features between clean transcripts and ASR transcripts to improve performance. In this paper, we first improve the ASR robustness of SLU tasks from a prototype perspective, and theoretically demonstrate the capability of our method to calibrate bias and shield partially bounded noise.

Contrastive Learning. As a promising paradigm of unsupervised learning, contrastive learning aims to maximize the similarities of positive pairs and

minimize negative pairs in a latent space (Arora et al., 2019; He et al., 2019). Contrastive learning is proposed for the first time in the field of computer vision and has achieved excellent performance in visual unsupervised representation (Chopra et al., 2005; Schroff et al., 2015; Chen et al., 2020; He et al., 2019; Zhu et al., 2024b; Chen et al., 2024). Recent studies introduce contrastive learning into the field of natural language processing and have achieved excellent performance in tasks such as sentence representation (Giorgi et al., 2020; Yan et al., 2021), visual-text modality (Radford et al., 2021), and so on. Chang and Chen (2022) and Cheng et al. (2023a) adopt contrastive learning to improve ASR robustness in SLU tasks. In this paper, we introduce contrastive clustering to achieve error calibration in SLU tasks based on prototypes.

Prototype Learning. Prototype learning (Wu et al., 2018) aims to learn prototypes as feature clustering centers from input representation belonging to the same category. Prototypical methods have been widely adopted in the field of computer vision (Snell et al., 2017; Wu et al., 2018; Wang et al., 2022). Pahde et al. (2020); Ji et al. (2020); Li et al. (2021) employ it to extract the overall characteristics of categories as a feature of imbalanced samples and apply it to few-shot learning. Jain et al. (2022); Kushwaha and Fuentes (2023) introduce the idea of prototypes to improve the performance of sound recognition. Wang et al. (2022) proposes a prototype-based method to calibrate erroneous pixels for semantic segmentation. Mustafa et al. (2020) proposes a prototype conformity loss to achieve high-quality adversarial defense. Li et al. (2020) utilizes prototypes as implicit variables and proposes prototypical contrastive learning to handle higher-level characteristics. Benefiting from these works, we focus on misclassified instance samples under noisy SLU tasks and introduce a prototype calibration loss with the capability of bias correction to improve ASR robustness.

Decoupling Training. Our method adopts teacher-student models to achieve decoupling training of clean transcripts and noisy ASR transcripts. Teacher-student models (Hinton et al., 2015) typically extract and compress knowledge through distillation, which has been successfully applied to a wide range of tasks such as semi-supervised learning and model compression. Tang et al. (2021) explores a teacher-student dual model framework for semi-supervised object detection. Cheng et al. (2023a) utilizes mutual learning to symmetrically train teacher and student models for ASR-robust SLU. In this paper, we introduce the idea of decoupling training to train clean and ASR transcripts. Unlike previous methods, we propose a novel EMA algorithm based on gradient sensitivity to achieve adaptive balance training for updating SLU models.

3 Method

In this section, we will elaborate on the specific pipeline of our PCAD. As shown in Figure 2, our proposed PCAD includes three components: (1) prototype calibration loss in pre-training (in section 3.2); (2) error-sensitive calibration loss in fine-tuning (in section 3.3); (3) adaptive-balanced decoupling training in fine-tuning (in section 3.4). Further, we provide the total training objective (in section 3.5) and the theoretical analysis (in section 3.6) of our method in ASR robustness.

3.1 Baselines based on Pre-training

Pre-training methods attempt to learn noise invariance and improve ASR robustness by contrastive pre-training on ASR transcripts and manual clean transcripts. Therefore, pre-training-based methods (Chang and Chen, 2022; Cheng et al., 2023b,a) usually consist of two stages: (1) adopting self-supervised contrastive learning to learn the invariant characteristics of noise between ASR transcripts and clean transcripts, and then (2) fine-tuning the pre-trained model in downstream tasks. These methods use teacher and student models to process ASR transcripts and clean transcripts respectively, and fine-tune them via distillation loss and cross-entropy loss. We follow two-stage settings and then use a prototype-based strategy to improve the robustness of these baselines.

3.2 Pre-training with Prototype Calibration

Given a mini-batch of input data of N pairs of texts $\mathcal{B} = \{(x_i^{clean}, x_i^{asr})\}_{i=1}^{N}$, we can obtain corresponding implicit states $\mathcal{E} = \{(e_i^p, e_i^q)\}_{i=1}^{N}$ through the last layer of `[CLS]` in pre-trained RoBERTa (Liu et al., 2019). Following Chang and Chen (2022) and Cheng et al. (2023a), we first utilize self-supervised contrastive learning to preliminarily align clean and noisy features:

$$\mathcal{L}^s = -\frac{1}{2N} \sum_{(e,e^+)\in\mathcal{P}} \log \frac{e^{\mathrm{s}(e,e^+)/\tau_s}}{\sum_{e\neq e}^{\mathcal{E}} e^{\mathrm{s}(e,e^+)/\tau_s}}, \quad (1)$$

where τ_s is the temperature coefficient, $\mathrm{s}(\cdot,\cdot)$ denotes the cosine similarity function and set $\mathcal{P} =$

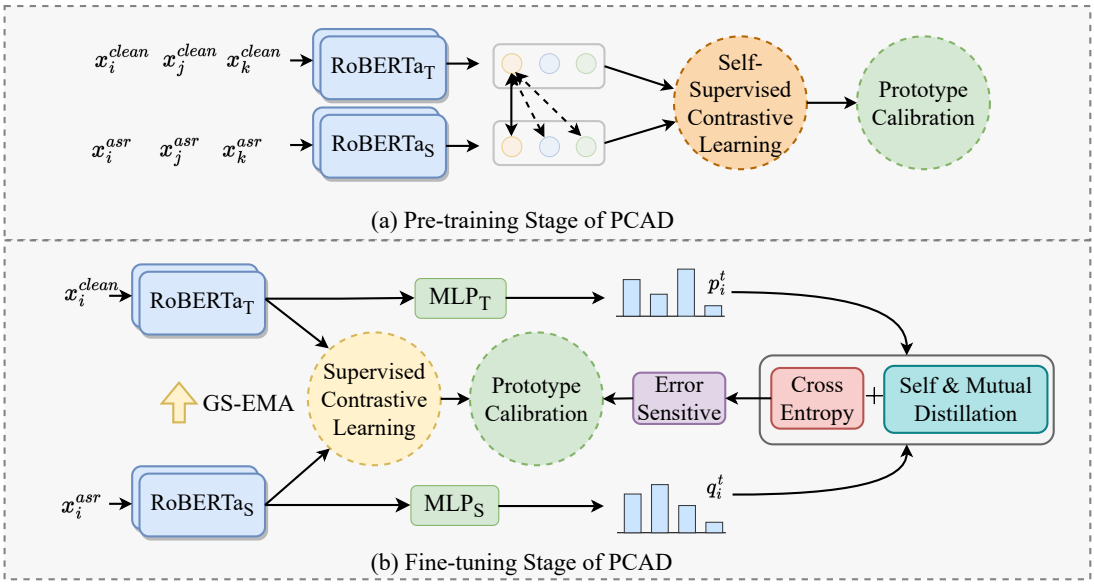


Figure 2: Overview of our PCAD. We illustrate the whole pipeline of the pre-training stage using PCL in (a) and the fine-tuning stage using ES-PCL and GS-EMA in (b).

$\{(e_i^p, e_i^q)\} \bigcup \{(e_i^q, e_i^p)\}$ as the matching set of aggregations (e_i^p, e_i^q) and (e_i^q, e_i^p).

Prototype Calibration Loss We adopt historical correct predictions to obtain prototype $\boldsymbol{\mu}_k \in \mathbb{R}^{1\times D}$ of category k. $\boldsymbol{\mu}_k$ is updated via exponential moving average (EMA) (Ye et al., 2019; Wu et al., 2018):

$$\boldsymbol{\mu}_k = \rho \boldsymbol{\mu}_k + (1-\rho)\frac{\sum_{i=1}^{|\mathcal{E}|}\delta(y_i=k)\cdot e_i}{1+\sum_{i=1}^{|\mathcal{E}|}\delta(y_i=k)}, \quad (2)$$

where ρ is the learning rate to update the prototypes and $\delta(\cdot)$ is the delta function. For convenience, we define $\mathcal{S}_k$ is a set of samples belonging to class k in $\mathcal{E}$, i.e., $\mathcal{S}_k = \{e_i \,|e_i \in \mathcal{E} \;\&\; y_i = k\}$.

We further define $s_{ik} = \tilde{\boldsymbol{\mu}}_k \tilde{e}_i$ as the similarity between the normalized vectors $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{e}_i$ in the unit hypersphere space, i.e., $\tilde{\boldsymbol{\mu}}_k = \frac{\boldsymbol{\mu}_k}{||\boldsymbol{\mu}_k||_2}$ and $\tilde{e}_k = \frac{e_k}{||e_k||_2}$. Therefore, the prototype calibration loss (PCL) with temperature τ_p used to alleviate sampling bias is represented as:

$$\mathcal{L}_i^p = -\log \frac{e^{s_{ik}/\tau_p}}{e^{s_{ik}/\tau_p} + \sum_{l\neq k} e^{s_{il}/\tau_p}}. \quad (3)$$

PCL calibrates the hyper-spherical representation of $\tilde{e}_i$ by bringing $\tilde{e}_i$ closer to $\tilde{\boldsymbol{\mu}}_k$ and pushing it apart from other prototypes, which injects rich label-semantic priors into the model.

3.3 Fine-tuning with Error-sensitive Prototype Calibration

In fine-tuning, we aggregate the features of manual and ASR transcripts into a set $\mathcal{S} = \{e_i\}_{i=1}^{2N}$. and we perform self-supervised contrastive learning (Khosla et al., 2020) in the same metric space:

$$\mathcal{L}^c = -\frac{1}{|\mathcal{S}|}\sum_{i}^{|\mathcal{S}|}\sum_{j\neq i}^{|\mathcal{S}|}\delta_{ij}\log\frac{e^{\mathrm{s}(e_i,e_j)/\tau_c}}{\sum_{k\neq i}^{|\mathcal{S}|}e^{\mathrm{s}(e_i,e_k)/\tau_c}}, \quad (4)$$

where y_i is the label of e_i, and τ is the temperature coefficient, $\delta_{ij} = \delta(y_i = y_j)$ is the delta function..

Error-sensitive Prototype Calibration Loss We approach correct and erroneous predictions with different strategies, making the model focus on explicit erroneous semantics. Following Wang et al. (2022), we firstly calculate the average cosine similarity between all misclassified samples and anchor e_i via the correction matrix:

$$\zeta_i = \begin{cases} \frac{1}{n_k^f}\sum_{j\in\mathcal{S}_k^f}(1-\mathrm{s}(\tilde{e}_i,\tilde{e}_j)) & if \;\; \mathcal{S}_k^f \neq \emptyset; \\ 0 & otherwise, \end{cases} \quad (5)$$

where $\mathcal{S}_k^f$ is a set of misclassified texts in $\mathcal{S}_k$, $n_k^f = |\mathcal{S}_k^f|$. The misclassified texts are pulled to the anchor e_i through minimizing ζ_i, thereby achieving remediation of erroneous intention predictions. In our work, ζ_i is a formal summary of the intention of misclassification. Motivated by Wang et al. (2022), we formulate the error-sensitive prototype calibration loss (ES-PCL) as:

$$\mathcal{L}_i^e = -\log\frac{e^{s_{ik}/\tau_e - \sigma(1-p_{ik})\zeta_i}}{e^{s_{ik}/\tau_e-\sigma(1-p_{ik})\zeta_i} + \sum_{l\neq k}e^{s_{il}/\tau_e}}, \quad (6)$$

where p_{ik} denotes the probability $P(y_i = k|e_i)$ obtained by the softmax operation; σ is the trade-off hyper-parameter and τ_e is the temperature.

3.4 Decoupling Learning for Fine-tuning

We train manual and ASR transcripts through teacher θ_t and student θ_s models respectively. Following Chang and Chen (2022); Cheng et al. (2023a), we adopt the strategies of self-distillation and mutual distillation to achieve bidirectional consistency in both time and model dimensions:

$$\begin{aligned}\mathcal{L}_t^d &= \frac{1}{N}\sum_i^N \mathrm{KL}_{\tau_d}(p_i^{t-1}||p_i^t),\\ \mathcal{L}_s^d &= \frac{1}{N}\sum_i^N \mathrm{KL}_{\tau_d}(q_i^{t-1}||q_i^t),\\ \mathcal{L}^m &= \frac{1}{N}\sum_i^N \mathrm{JS}_{\tau_m}(p_i^t||q_i^t),\end{aligned} \tag{7}$$

where p_i^t and q_i^t denote the probability output via teacher and student models at t-th stage; KL_{τ_d} and JS_{τ_m} are KL and JS divergence with different temperatures. Therefore, the total distillation loss is

$$\mathcal{L}^d = \mathcal{L}_t^d + \mathcal{L}_s^d + \mathcal{L}^m. \tag{8}$$

Adaptive Decoupling Training In order to achieve balanced training of the teacher-student model, we need to find a dynamic indicator to measure the sensitivity between the teacher and student branches during the training phase. Since teacher and student models are jointly optimized through gradient backpropagation, the magnitude of the gradients reflects the sensitivity of the two branches to input data. Motivated by this, we propose a novel gradient-sensitive EMA algorithm to update models in an asymmetric manner adaptively:

$$\theta_t = \rho_o\theta_t + (1-\rho_o)|\frac{S_p}{S_q}|\theta_s, \tag{9}$$

where ρ_o is the balanced hyper-parameter. We utilize the average L2-norm of gradients to measure the significance of the models θ_t and θ_s:

$$S_p = ||\frac{\partial \mathcal{L}_{ce}^p}{\partial e^p}||_2, \quad S_q = ||\frac{\partial \mathcal{L}_{ce}^q}{\partial e^q}||_2, \tag{10}$$

where $\mathcal{L}_{ce}^p$ and $\mathcal{L}_{ce}^q$ are the cross-entropy loss calculated by using p_i^t and q_i^t:

$$\mathcal{L}_{ce}^p = -\sum_{i=1}^N y_i^p \log p_i^t, \quad \mathcal{L}_{ce}^q = -\sum_{i=1}^N y_i^q \log q_i^t. \tag{11}$$

Then, we can obtain the total predicted loss as:

$$\mathcal{L}_{ce} = \mathcal{L}_{ce}^p + \mathcal{L}_{ce}^q. \tag{12}$$

3.5 Training Objective and Analysis

Pre-training Following Chang and Chen (2022); Cheng et al. (2023a), we adopt an MLM loss $\mathcal{L}_{mlm}$ as the regularization term. And the total pre-training loss is :

$$\mathcal{L}_{pt} = \mathcal{L}^s + \lambda\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\mathcal{L}_i^p + \eta\mathcal{L}^{mlm}, \tag{13}$$

where λ and η are the trade-off hyper-parameters.

Fine-tuning The complete fine-tuning loss is:

$$\mathcal{L}_{ft} = \mathcal{L}_{ce} + \alpha\mathcal{L}^c + \beta\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathcal{L}_i^e + \gamma\mathcal{L}^d, \tag{14}$$

where α, β and γ are the trade-off hyper-parameters.

3.6 Theoretical Analysis

We analyze the effectiveness of our method on ASR robustness from the perspectives of gradient and decision, respectively:

Gradient perspective Similar to other prototype learning methods (Chen et al., 2022; Sharma et al., 2023; Wang et al., 2022; Li et al., 2020), we analyze the effectiveness of $\mathcal{L}^p$ and $\mathcal{L}^e$ on bias and error calibration from a gradient perspective. Since $\mathcal{L}^p$ is a lower bound of $\mathcal{L}^e$, we take $\mathcal{L}^e$ as an example to elaborate the effectiveness of prototype calibration. We can calculate the gradient of $\mathcal{L}_i^e$ w.r.t feature e_i as follows:

$$\begin{aligned}\frac{\partial\mathcal{L}_i^e}{\partial e_i} = &\underbrace{(\frac{e^{s_{ik}/\tau_e}}{\sum_{l'}e^{s_{il'}/\tau_e}} - 1)\frac{\tilde{\boldsymbol{\mu}}_k}{\tau_e} + \sum_{l\neq k}\frac{e^{s_{il}/\tau_e}}{\sum_{l'}e^{s_{il'}/\tau_e}}\frac{\tilde{\boldsymbol{\mu}}_l}{\tau_e}}_{\text{bias calibration}}\\ &+\underbrace{(\frac{e^{s_{ik}/\tau_e}}{\sum_{l'}e^{s_{il'}/\tau_e}} - 1)\cdot(\frac{\sigma(1-p_{ik})}{n_k^f}\sum_{j\in\mathcal{S}_k^f}e_j)}_{\text{error calibration}}.\end{aligned} \tag{15}$$

We can observe that $\partial\mathcal{L}_i^e/\partial e_i$ can be formulated into two terms: **(1)** The bias calibration achieves intra- and inter-class gradient compensation to alleviate feature deterioration caused by sampling bias. This item directly calibrates the gradient by making e_i close the intra-class prototype $\tilde{\boldsymbol{\mu}}_k$ and away

from inter-class prototypes $\tilde{\boldsymbol{\mu}}_l$. **(2)** The error calibration provides gradient penalties for erroneous predictions set $\mathcal{S}_k^f$. This item adds a penalty for incorrect predictions e_j, making the weight of incorrect predictions greater than that of the correct predictions.

Decision Perspective Following the analysis of convex outer adversarial polytope (Kolter and Wong, 2017), we first define the representation of polytope space and the average spatial overlap coefficient as follows:

Definition 1. *We mathematically define a representation polytope space under an ASR noise ϵ with p norm bounded by δ for input samples x through the pre-trained model RoBERTa θ :*

$$\mathcal{S}_\delta(x;\theta) = \{\text{RoBERTa}(x+\epsilon)\ s.t., ||\epsilon||_p \leq \delta\}. \tag{16}$$

Definition 2. *The space overlap between two samples x_i and x_j belonging to different categories y_i and y_j can be defined as the set of intersections between the respective polytope space:*

$$\mathcal{SO}_\delta(x_i, x_j) = \mathcal{S}_\delta(x_i;\theta) \bigcap \mathcal{S}_\delta(x_j;\theta). \tag{17}$$

Then, the averaged spatial overlap coefficients for all samples can be defined as:

$$\mathcal{ASO}_\delta = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, y_j \neq y_i}^{M} \mathcal{SO}_\delta(x_i, x_j), \tag{18}$$

where M denotes the number of all samples.

Based on this, we argue that our prototype loss can improve robustness as follows:

Claim 1. ***(The existence of robust classifiers.)*** *When we employ a classifier that maintains a margin m between the two closest samples belonging to different classes:*

$$m > \max_{x,y \in \mathcal{S}_\delta(x_i;\theta)} d(x, y), \tag{19}$$

where $d(\cdot,\cdot)$ denotes the distance between x and y, we can obtain a decision boundary with guaranteed robustness against a bounded noisy ϵ.

Claim 2. ***(ES-PCL improves robustness by reducing $\mathcal{ASO}_\delta$.)*** *Based on Claim 1, we can obtain a classifier that is robust to ASR errors by changing the spatial overlap coefficient of the convex polytope. Our ES-PCL in Eq. 6 can reduce the average spatial overlap coefficient $\mathcal{ASO}_\delta$ by pushing out inter-class prototypes and pulling intra-class prototypes, thereby improving the robustness to ASR noise on the whole dataset.*

Claims 1 and 2 reveal that the robustness of the model is directly related to the overlap coefficient of the representation space between the perturbed samples and other categories of samples. Our prototype-based method implicitly obtains a convex class-specific classification area for each category and uses prototypes as the center of polytope space to reduce overlap volume. Therefore, our method is more robust to bounded perturbations of samples while ensuring classification accuracy.

4 Experiment

Datasets and Metric Following previous work (Chang and Chen, 2022; Cheng et al., 2023b,a), we conduct all the experiments on three publicly available benchmark datasets[1]: SLURP (Bastianelli et al., 2020; Chang and Chen, 2022), ATIS (Hemphill et al., 1990) and TREC6 (Li and Roth, 2002). The statistics of the three datasets included are shown in Table 2. SLURP consists of several (scenario, action) pairs and the prediction is considered correct only when both the scenario and action are predicted correctly. ATIS and TREC6 are two public SLU datasets for flight reservation and question classification, respectively. Similar to (Chang and Chen, 2022; Cheng et al., 2023a), we utilize the synthesized text provided by Phoneme-BERT (Sundararaman et al., 2021). The ASR transcripts are obtained by Google Web API. We use accuracy as the metric on ATIS and TREC6 and joint accuracy on SLURP for intent classification.

Training Settings and Baselines For fair comparison with (Chang and Chen, 2022; Cheng et al., 2023a), we pre-train PCAD for $10K$ steps with a batch size 128 on each dataset, and finetune the whole model up to 10 epochs with a batch size of 256. We adopt a early-stop strategy and utilize Adam (Kingma and Ba, 2015) as optimizer with β_1 = 0.9, β_2= 0.98. The mask ratio of MLM is set to 0.15. The temperature coefficients τ_s, τ_p, τ_c, τ_p, τ_d and τ_m are set to 0.2, 0.1, 0.2, 0.1, 5 and 1, respectively. The hyper-parameter σ in Eq. 6 is set to 0.5. ρ in Eq. 2 is set to 0.99. λ and η in Eq. 13 are empirically set to 0.5 and 1. α, β, and γ in Eq. 14 are empirically set to 0.5, 0.5 and 1, respectively. The balanced hyper-parameter ρ_o is set to 0.99. All reported accuracy scores of our method in this paper are averaged over 5 runs. All experiments are

[1]SLURP is available at https://github.com/MiuLab/SpokenCSE, and ATIS and TREC6 are available at https://github.com/Observeai-Research/Phoneme-BERT.

Methods	w/o manual transcripts in fine-tuning		
	SLURP	ATIS	TREC6
RoBERTa (Liu et al., 2019)	83.97	94.53	84.08
Phoneme-BERT (Sundararaman et al., 2021)	83.78	94.83	85.96
SimCSE (Gao et al., 2021)	84.47	94.07	84.92
SpokenCSE (Chang and Chen, 2022)	85.26	95.10	86.36
ML-LMCL (Cheng et al., 2023a)	88.52	96.52	89.24
PCAD w/o PCL	89.15 (↓1.43)	96.78 (↓0.86)	89.99 (↓1.26)
PCAD w/o ES-PCL	88.76 (↓1.82)	96.57 (↓1.07)	89.60 (↓1.65)
PCAD w/o GS-EMA	89.93 (↓0.65)	97.08 (↓0.56)	90.27 (↓0.98)
PCAD (Full)	**90.58***	**97.64***	**91.25***

Table 1: Results of intention classification accuracy on three datasets. $*$ means the improvement of PCAD is statistically significant with $p<0.05$ under t-test. "w/o manual transcripts" denotes manual transcripts are not used in fine-tuning, and ASR transcripts are fed into both teacher and student models.

Dataset	#Class	Avg. Length	Train	Test
SLURP	18×46	6.93	50,628	10,992
ATIS	22	11.14	4,978	893
TREC6	6	8.89	5,452	500

Table 2: The statistics of the three SLU datasets. The test set of SLURP dataset is sub-sampled following Chang and Chen (2022) and Cheng et al. (2023a).

conducted on 8 NVIDIA RTX3090 GPUs. We compare our model with five state-of-the-art baselines, including RoBERTa (Liu et al., 2019), Phoneme-BERT (Sundararaman et al., 2021), SimCSE (Gao et al., 2021), SpokenCSE (Chang and Chen, 2022) and ML-LMCL (Cheng et al., 2023a).

4.1 Main Results

We conduct a series of quantitative experiments on three datasets, i.e., SLURP ATIS and TREC6, to compare our method with baselines. Table 1 and Table 3 show the experimental results of different models without and with clean transcripts during the fine-tuning stage, respectively. From the results, we have the following observations:

(1) Our PCAD (Full) consistently outperforms all baselines on all tasks and datasets. Without manual transcripts during fine-tuning, our PCAD achieves significant improvements in the accuracy of 2.06%, 1.12%, and 2.01% on SLURP, ATIS, and TREC6 compared to the best baseline (*i.e.*, ML-LMCL), even surpassing the ML-LMCL with clean transcripts in fine-tuning. PCAD also achieves significant improvements with manual transcripts in fine-tuning, demonstrating the effectiveness of our calibration and decoupling strategies.

(2) Our PCAD (Full) achieves significant performance improvement on the SLURP dataset. The task on SLURP is more challenging, which requires the correct classification of scenarios and action intentions simultaneously. Our method has greatly improved the joint accuracy of intention pairs, achieving the best joint accuracy (90.58% and 91.89%) on SLURP without and with manual transcripts in fine-tuning. The result further indicates the significant advantages of our method in improving accuracy and robustness in SLU tasks with ASR noise.

4.2 Components Analysis

Effects of the prototype-based loss One of the core contributions of our PCAD is the prototype-based calibration loss, which alleviates bias during the pre-training phase and reduces erroneous predictions in fine-tuning. As shown in Table 1 and Table 3, we conduct a series of experiments to evaluate the effectiveness of PCL and ES-PCL. It can be observed that PCL and ES-PCL contribute to the performance positively during pre-training and fine-tuning, respectively. Removing any component results in a significant decrease in model performance, e.g., reducing joint accuracy by 1.43% and 1.82% on SLURP in Table 1 without PCL and ES-PCL, respectively. The results show that our calibration strategy for pre-training and fine-tuning can achieve dual improvements in the accuracy and robustness of SLU with ASR noise.

Effects of GS-EMA To verify the effectiveness of adaptive-balanced decoupling learning, we remove the GS-EMA strategy to conduct experi-

Methods	w/ manual transcripts in fine-tuning		
	SLURP	ATIS	TREC6
RoBERTa (Liu et al., 2019)	84.42	94.86	84.54
Phoneme-BERT (Sundararaman et al., 2021)	84.16	95.14	86.48
SimCSE (Gao et al., 2021)	84.88	94.32	85.46
SpokenCSE (Chang and Chen, 2022)	85.64	95.58	86.82
ML-LMCL (Cheng et al., 2023a)	89.16	97.21	89.96
PCAD w/o PCL	90.39 (↓1.50)	97.27 (↓0.78)	91.03 (↓1.30)
PCAD w/o ES-PCL	90.14 (↓1.75)	96.95 (↓1.10)	90.45 (↓1.88)
PCAD w/o GS-EMA	90.20 (↓1.69)	97.24 (↓0.81)	90.80 (↓1.53)
PCAD (Full)	**91.89***	**98.05***	**92.33***

Table 3: Results on three datasets. * means the improvement of PCAD is statistically significant with $p<0.05$ under t-test. "w manual transcripts" denotes manual transcripts are used in fine-tuning. Clean manual transcripts and ASR transcripts are fed into teacher and student models, respectively.

Strategy	SLURP	ATIS	TREC6
w/o EMA	90.20	97.24	90.80
Vanilla EMA	90.72	97.61	91.55
GS-EMA	**91.89**	**98.05**	**92.33**

Table 4: Ablation study of different EMA methods.

Prototype	ES	SLURP	ATIS	TREC6
Learnable	×	90.44	97.15	90.67
All predictions	×	90.53	97.29	90.91
Correct	×	90.97	97.56	91.30
Correct	✓	**91.89**	**98.05**	**92.33**

Table 5: Ablation experiments on different types of prototypes. ES denotes the error-sensitive regularizer in Eq. 5. "Correct" means only correct predictions are used to update prototypes.

ments as shown in Table 1 and Table 3. We can observe that when there is no clean transcript for fine-tuning (i.e., ASR hypotheses are fed into both teacher and student branches), accuracy drops by 0.65%, 0.56% and 0.98% on SLURP, ATIS and TREC6, respectively. In addition, the contribution of GS-EMA components to performance is significantly improved when both clean and noisy transcripts are available, i.e., accuracy drops by 1.69%, 0.81% and 1.53% on SLURP, ATIS and TREC6. The results prove that the GS-EMA strategy can indeed promote the teacher model to effectively extract knowledge from the student model to enhance ASR noise robustness. We further replace GS-EMA with a regular EMA algorithm for experiments, as shown in Table 4. We can observe that GS-EMA achieves more significant performance than the vanilla EMA, which proves that our adaptive balancing algorithm based on gradient sensitivity is effective in improving the performance under noisy settings.

Analysis of the prototype As shown in Table 5, we conduct an experiment to evaluate the effects of different types of prototypes. We use learnable parameters, all predictions, and correct predictions as prototypes, respectively. We observe that the best accuracy can be obtained when we leverage correct predictions as prototypes. We suppose the reason for this is that correct predictions avoid injecting erroneous semantics compared to other prototypes. In addition, the learnable prototypes use a set of learnable parameters as prototypes, which ignores category label information and obtains the lowest performance.

4.3 Robustness analysis

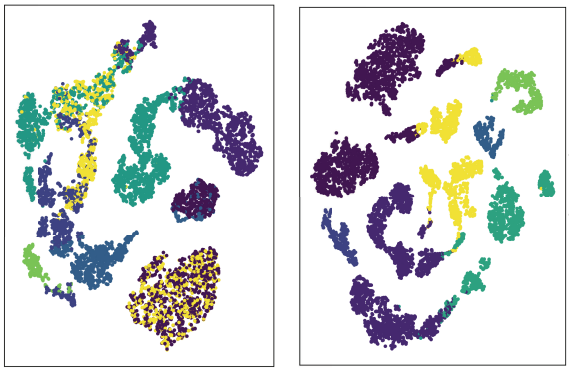
Following Chang and Chen (2022), we divide the test set into four subsets based on the word error rate (WER) of ASR transcripts to evaluate the robustness of our method under different noise levels in Table 6. It can be observed that: (1) Our method outperforms all previous methods under all noise levels on SLURP, i.e., under different WER intervals; (2) When inputting clean transcripts (i.e., WER=0), our model improves by 0.68% in accuracy compared to SpokenCSE, which indicates that PCAD also has great potential in the original SLU task without ASR errors; (3) As the noise level increases, the performance improvement achieved by our model becomes more significant. In the case of the highest noise, our method improves performance by 5.13% in accuracy compared to SpokenCSE. This verifies that our strategy has re-

Method	clean 0	low WER (0,0.16]	medium WER (0.16,0.4]	high WER > 0.4
RoBERTa (Liu et al., 2019)	95.69	92.41	85.89	56.71
Phoneme-BERT (Sundararaman et al., 2021)	94.97	92.34	85.87	57.20
SimCSE (Gao et al., 2021)	95.55	93.47	86.82	57.59
SpokenCSE (Chang and Chen, 2022)	96.08	94.41	87.63	58.72
PCAD (Full)	**96.76**	**96.65**	**90.92**	**63.85**

Table 6: Experimental results under different WER intervals on SLURP dataset.

silience to higher ASR noise and can significantly improve the ASR robustness of the model.

4.4 Qualitative analysis



(a) SpokenCSE (b) Our PCAD

Figure 3: Feature visualization of the SLURP test set under different methods (in color). We show the 2D point map of the intermediate features obtained from SpokenCSE and our PCAD, where clusters with different colors represent different categories.

To exhibit the improvement of intra-class consistency and inter-class discrimination of the feature representation by our PCAD, we randomly select 4096 samples on the SLURP test set and visualize their intermediate features on SpokenCSE and our PCAD through the T-SNE toolkit (van der Maaten and Hinton, 2008). As shown in Figure 3, we can observe that PCAD can better achieve the separation of inter-class features and the aggregation of intra-class features, avoiding feature aliasing.

5 Conclusion

In this paper, we introduce prototype-based losses for pre-training and fine-tuning, namely PCL and ES-PCL, and analyze their remedial effects on sampling bias and semantic errors from a theoretical perspective. It is the first time to improve ASR robustness from the perspective of prototype calibration. In addition, in order to achieve high-quality decoupling training, we propose a novel update algorithm based on gradient saliency for adaptive balance. Extensive experiments on three datasets demonstrate that PCAD outperforms the state-of-the-art approaches by a large margin.

Limitations

This article proposes a framework to improve the robustness of SLU models in practical scenarios. Our framework can accurately calibrate ASR errors and improve the performance of pipeline SLU models in practical scenarios. However, the training of our proposed model relies on aligning clean manual transcripts with ASR transcripts, which may not be easily obtained in specific scenarios. Obtaining clean artificial transcripts is time-consuming and labor-intensive. Therefore, exploring how to reduce the proportion of manual clean transcripts or train models only using ASR noise transcripts is more valuable for practical scenarios, which also is an important direction for future work.

Ethics Statement

We conduct all experiments on the public datasets, which do not contain any offensive content or information with negative social impact. The focus of our paper is to improve the noise robustness of the SLU model and our model does not have uncontrollable outputs. Therefore, we ensure that our paper complies with ethical review guidelines.

Acknowledgements

We would like to thank all reviewers for their insightful comments. This paper was partially supported by NSFC (No: 62176008). Special acknowledgements are given to AOTO-PKUSZ Joint Research Center for its support.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A Spoken Language Understanding Resource Package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yanfeng Chang and Yun-Nung Chen. 2022. Contrastive learning for improving asr robustness in spoken language understanding. In *Interspeech*.

Cheng Chen, Ji Zhang, Jingkuan Song, and Lianli Gao. 2022. Class gradient projection for continual learning. *Proceedings of the 30th ACM International Conference on Multimedia*.

Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, and Masashi Sugiyama. 2021. Large-margin contrastive learning with distance polarization regularizer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1673–1683. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505.

Xuxin Cheng, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023b. C²A-SLU: Cross and Contrastive Attention for Improving ASR Robustness in Spoken Language Understanding. *INTERSPEECH 2023*.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, Xianwei Zhuang, and Yuexian Zou. 2024. Towards multi-intent spoken language understanding via hierarchical attention and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 17844–17852.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

John Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *ArXiv*, abs/2006.03659.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The atis spoken language systems pilot corpus. In *Human Language Technology - The Baltic Perspectiv*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Chao-Wei Huang and Yun-Nung (Vivian) Chen. 2019. Adapting pretrained transformer to lattices for spoken language understanding. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 845–852.

Dhruv Jain, Khoa Nguyen, Steven M. Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon E. Froehlich. 2022. Protosound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Zhong Ji, Xingliang Chai, Yunlong Yu, Yanwei Pang, and Zhongfei Zhang. 2020. Improved prototypical networks for few-shot learning. *Pattern Recognit. Lett.*, 140:81–87.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *ArXiv*, abs/2004.11362.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

J. Zico Kolter and Eric Wong. 2017. Provable defenses against adversarial examples via the convex outer adversarial polytope. *ArXiv*, abs/1711.00851.

Saksham Singh Kushwaha and Magdalena Fuentes. 2023. A multimodal prototypical approach for unsupervised sound classification. *ArXiv*, abs/2306.12300.

Gen Li, V. Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. 2021. Adaptive prototype learning and allocation for few-shot segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8330–8339.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. 2020. Prototypical contrastive learning of unsupervised representations. *ArXiv*, abs/2005.04966.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *International Conference on Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. Asr error correction and domain adaptation using machine translation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348.

Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. In *NIPS*.

Aamir Mustafa, Salman Hameed Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. 2020. Deeply supervised discriminative learning for adversarial defense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3154–3166.

Frederik Pahde, Mihai Marian Puscas, Tassilo Klein, and Moin Nabi. 2020. Multimodal prototypical networks for few-shot learning. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2643–2652.

Martin H. Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann. 2020. End-to-end neural transformer based spoken language understanding. In *Interspeech*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Saurabh Sharma, Yongqin Xian, Ning Yu, and Ambuj K. Singh. 2023. Learning prototype classifiers for long-tailed recognition. *ArXiv*, abs/2302.00491.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *NIPS*.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript. In *Interspeech*.

Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. 2021. Humble teachers teach better students for semi-supervised object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140.

Gokhan Tur and Renato De Mori. 2011. Spoken language understanding: Systems for extracting semantic information from speech. In .

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Haoyu Wang, Shuyan Dong, Yue Liu, James Logan, Ashish Kumar Agrawal, and Yang Liu. 2020. Asr error correction with augmented transformer for entity retrieval. In *Interspeech*.

Hualiang Wang, Huanpeng Chu, Siming Fu, Zuozhu Liu, and Haoji Hu. 2022. Renovate yourself: Calibrating feature representation of misclassified pixels for semantic segmentation. In *AAAI Conference on Artificial Intelligence*.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *ArXiv*, abs/2105.11741.

Mang Ye, Xu Zhang, PongChi Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6203–6212.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Annual Meeting of the Association for Computational Linguistics*.

Linchen Zhu, Wenjie Liu, Linquan Liu, and Ed Lin. 2021. Improving asr error correction using n-best hypotheses. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 83–89.

Zhihong Zhu, Xuxin Cheng, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2024a. Aligner²: Enhancing joint multiple intent detection and slot filling via adjustive and forced cross-task alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19777–19785.

Zhihong Zhu, Xianwei Zhuang, yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024b. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024. Towards explainable joint models via information theory for multiple intent detection and slot filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17, pages 19786–19794.