RUCL: Integrating Regularization and Unsupervised Contrastive Learning for Automatic Speech Recognition

Zeting Li*, Jiangping Zhu*†, Pei Zhou*†, Zhengzhong Zhu†
*Nat. Key Lab. of Fund. Sci. on Synthetic Vision, Sichuan University, Chengdu, China
†College of Computer Science, Sichuan University, Chengdu 610065, China
Corresponding author: zjp16@scu.edu.cn

***Abstract*—Automatic Speech Recognition (ASR) aims to transcribe speech signals into textual sequences, serving as a fundamental technology in human-computer interaction. However, end-to-end ASR systems often suffer from inconsistent predictions across different stochastic passes and insufficient representation learning capability. To mitigate these issues, we propose RUCL (Regularization and Unsupervised Contrastive Learning), a novel framework that hierarchically integrates consistency regularization and unsupervised contrastive learning.Specifically, at the feature level, we employ unsupervised contrastive learning to enhance the discriminability of speech representations, thereby strengthening the model's ability to capture robust acoustic features. At the output level, we apply consistency regularization to minimize the divergence between two CTC distributions generated under different dropout masks, promoting prediction stability. Extensive experiments on AISHELL-1 and LibriSpeech datasets demonstrate that RUCL achieves superior performance, with a CER of 4.63% on AISHELL-1 and WERs of 4.44%/9.96% on LibriSpeech test-clean/test-other, significantly outperforming strong baselines.**

***Index Terms*—Automatic Speech Recognition, Consistency Regularization, Unsupervised Contrastive Learning.**

I. Introduction

Automatic Speech Recognition (ASR) seeks to transcribe spoken utterances into textual form. End-to-end approaches [1]–[3], which directly model the mapping from acoustic features to text sequences, have become the dominant paradigm due to their architectural simplicity and training efficiency. Prominent among these are methods based on Connectionist Temporal Classification (CTC) [1], the Attention-based Encoder-Decoder (AED) [2], and hybrid CTC/AED frameworks [4], which leverage complementary advantages to enhance recognition accuracy.

Despite progress in architecture, current end-to-end ASR systems still face a fundamental challenge in representation learning. Mainstream methods rely on sparse and frame-level ambiguous supervision from CTC or attention alignment [1], leading the encoder to capture shallow acoustic patterns rather than deep discriminative features. This causes poor inter-class separation for similar phonemes and low intra-class consistency under noise. Weak representations further exacerbate prediction inconsistency: under stochastic perturbations such as dropout, the encoder produces high-variance outputs with unreasonable probability "spikes", reducing robustness to local noise interference [5].

To mitigate these issues, prior work has explored various regularization strategies. Consistency regularization stabilizes predictions by enforcing invariance across different views. Methods such as R-Drop [5] and Yao et al. [6] minimize divergence between output distributions, while other works reduce feature redundancy [7]. Although effective in smoothing outputs, these methods mainly constrain final predictions and do not explicitly enhance the discriminability of low-level speech representations.

Another direction is Self-supervised Speech Representation Learning (SSL), which uses unlabeled data to learn robust features. Frameworks like Wav2vec 2.0 [8] and HuBERT [9] achieve remarkable success via contrastive learning or masked prediction. However, they involve high costs due to large-scale pre-training and rely on complex external processes, with objectives often decoupled from the ASR task. Recently, multi-view mining and unsupervised semantic alignment have shown potential in improving representation discriminability [10], [11].

We thus ask: Can we design a unified, lightweight framework that enhances feature discriminability and output stability during standard training, without external augmentation or large-scale pre-training? To this end, we propose RUCL, a novel ASR framework that hierarchically integrates consistency regularization and unsupervised contrastive learning in a Conformer-Transformer hybrid CTC/AED architecture. The model processes input speech twice with different dropout masks to obtain two sets of encoded features and CTC distributions. At the feature level, unsupervised contrastive learning pulls representations of the same utterance closer and pushes others apart. At the output level, consistency regularization minimizes distribution divergence. The overall objective combines standard losses with the two auxiliary losses. Unlike existing works, RUCL uses intrinsic dropout noise to construct multi-view constraints in a unified framework.

Our main contributions are summarized as follows:

- We propose RUCL, an ASR framework that hierarchically integrates consistency regularization and unsupervised contrastive learning to enhance discriminability and stabilize CTC outputs.

- RUCL leverages intrinsic dropout noise to construct multi-view features for joint learning, without external data or model modification.
- Experiments on AISHELL-1 and LibriSpeech show that RUCL significantly improves recognition performance, surpassing strong baselines including self-supervised models and state-of-the-art regularization methods [7].

II. Related work

A. Consistency Regularization for Robust ASR Prediction

Consistency regularization enhances robustness by enforcing invariant predictions across different stochastic passes or augmented views. Originating from visual representation learning [12] and semi-supervised learning [13], this strategy minimizes divergence between perturbed representations to avoid overfitting. In speech processing, it is applied via reconstruction objectives [14] or spatial-temporal constraints [15], often combined with SpecAugment [16] for view diversity. In supervised ASR, typical methods include R-Drop [5], which uses bidirectional KL-divergence to stabilize outputs, and Yao et al. [6] with Zipformer [17]. Ko et al. [7] further proposed diversity regularization to reduce feature redundancy. These methods mitigate prediction inconsistency but mainly focus on output stability, without explicitly improving the discriminability of encoder representations.

B. Contrastive Learning for Discriminative Speech Representation

Contrastive learning learns discriminative representations by pulling positive pairs closer and pushing negative pairs apart in the latent space, directly addressing insufficient feature discriminability. In speech processing, self-supervised methods such as Wav2Vec 2.0 [8] and HuBERT [9] achieve great success. Wav2Vec 2.0 uses contrastive loss to distinguish real quantized speech representations, while HuBERT performs masked prediction with clustered hidden units as pseudo-labels. Both enhance the ability to capture robust content-sensitive acoustic features from unlabeled data.

Despite effectiveness, these methods usually need large-scale pre-training and rely on complex data augmentation or offline clustering to build positive pairs. In contrast, RUCL constructs natural positive pairs without external augmentation. It improves encoder discriminability while compatible with standard CTC and attention losses, providing a unified way to boost feature quality and robustness.

C. Peak Suppression in CTC-based ASR

A well-known symptom of prediction inconsistency in CTC-based models is producing highly "peaky" distributions, where non-blank tokens appear as sharp probability spikes [18]. This indicates overfitting to training alignments and causes issues like lower boundary precision in forced alignment [19] and knowledge distillation difficulties [20]. Peak suppression aims to encourage smoother and more reasonable probability distributions.

To address this, existing methods use label priors [19] or sequence-level knowledge distillation [21]. They penalize overconfident predictions by modifying training objectives or decoding, but work as separate adjustments rather than being integrated into representation learning. Our RUCL provides a synergistic solution that links peak suppression with representation learning. By enforcing consistency across stochastic outputs, RUCL suppresses sharp spikes and yields smoother, more robust distributions.

III. Method

The proposed hierarchical consistency regularization framework operates at two levels: (a) feature-level consistency: pulling closer the feature representations obtained from the same input under different dropout masks; and (b) output-level consistency, which enforces stability in the output probability distributions across these views. These two components synergize to form a hierarchical constraint system spanning from low-level features to high-level outputs. The model structure is illustrated in Fig. 1.

A. Consistency Regularization

To mitigate the issue of inconsistent predictions in CTC-based ASR models under different stochastic passes, we introduce a consistency regularization method inspired by R-Drop [5]. The core idea is to minimize the divergence between two output distributions generated from the same input under different dropout conditions. This effectively penalizes the sharpness of the loss landscape, guiding the optimization toward a flatter minimum.

Formally, let $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}$ denote the input speech feature sequence. We pass $\mathbf{X}$ through the shared Conformer encoder twice, each with a distinct dropout mask $\boldsymbol{\xi}^{(a)}$ and $\boldsymbol{\xi}^{(b)}$, yielding two sets of hidden representations:

$$\mathbf{Z}^{(a)} = f^1(\mathbf{X}; \boldsymbol{\xi}^{(a)}), \quad \mathbf{Z}^{(b)} = f^2(\mathbf{X}; \boldsymbol{\xi}^{(b)}), \tag{1}$$

where $\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)} \in \mathbb{R}^{B \times T \times D}$ with B as the batch size, T as the sequence length, and D as the feature dimension. These representations are then processed by a shared CTC head to generate two probability distributions over the output vocabulary:

$$\begin{aligned} \mathbf{P}^{(a)} &= \mathrm{softmax}(\mathbf{W}\mathbf{Z}^{(a)} + \mathbf{b}), \\ \mathbf{P}^{(b)} &= \mathrm{softmax}(\mathbf{W}\mathbf{Z}^{(b)} + \mathbf{b}), \end{aligned} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{D \times |\mathcal{V}|}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$ apply a linear projection to the vocabulary size $|\mathcal{V}|$, followed by the softmax operation.

To enforce consistency between these two distributions, we minimize the bidirectional Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2}\left[\mathcal{D}_{\mathrm{KL}}(\mathbf{P}^{(a)} \| \mathbf{P}^{(b)}) + \mathcal{D}_{\mathrm{KL}}(\mathbf{P}^{(b)} \| \mathbf{P}^{(a)})\right], \tag{3}$$

This loss term encourages the model to produce similar output distributions regardless of the dropout noise, thereby stabilizing the CTC predictions and mitigating the inconsistency issue.

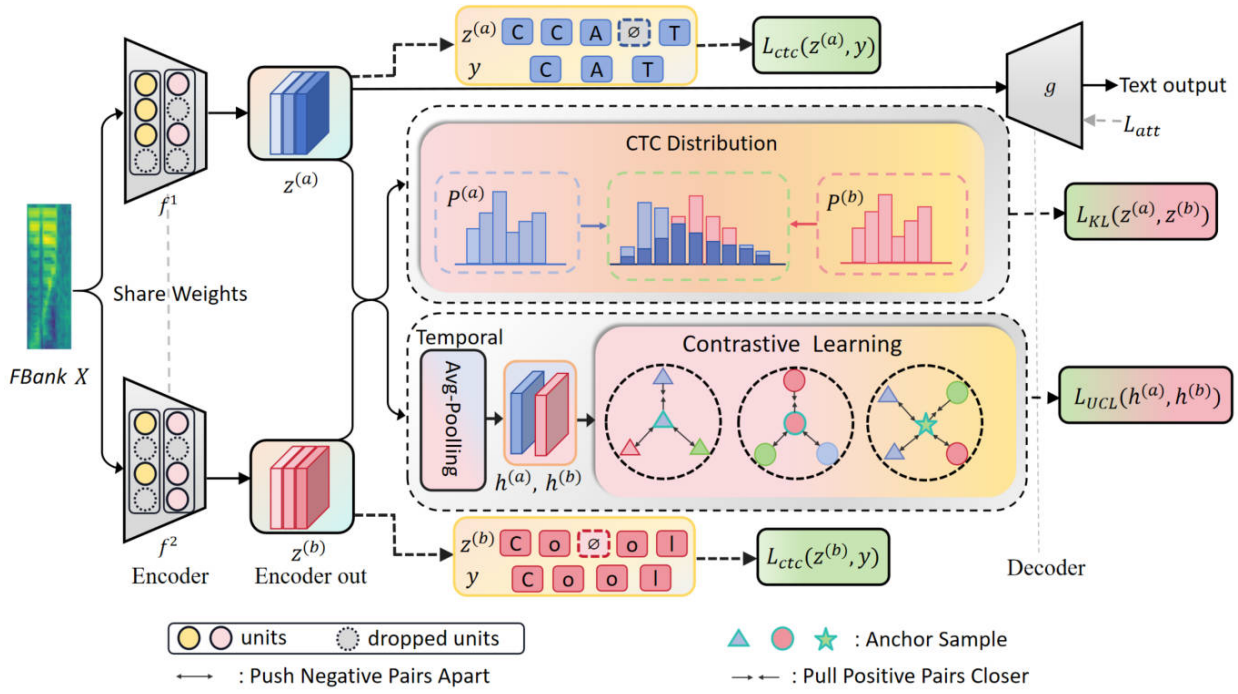


Fig. 1: Overview of the proposed RUCL framework. The input FBank features are processed through two forward passes of the same Conformer encoder with different dropout masks, yielding two sets of hidden representations $\mathbf{Z}^{(a)}$ and $\mathbf{Z}^{(b)}$. These representations are used to compute the CTC losses and the bidirectional KL-divergence loss for consistency regularization (Section III-A). Simultaneously, utterance-level features $\mathbf{h}^{(a)}$ and $\mathbf{h}^{(b)}$, obtained via temporal average pooling, are fed into a contrastive learning module to compute the unsupervised contrastive loss (Section III-B).

B. Unsupervised Contrastive Learning

To tackle the issue of insufficient discriminability in speech representations learned by the encoder, we introduce a RUCL module. This component operates at the feature level, leveraging the multi-view representations generated by dropout to enhance the encoder's ability to capture robust and discriminative acoustic features.

To obtain global semantic information and align with the requirements of instance-level contrastive learning, we apply temporal average pooling to the frame-level features $\mathbf{Z}^{(a)}$ and $\mathbf{Z}^{(b)}$ (derived in Section III-A) along the time dimension:

$$\mathbf{h}^{(a)} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_t^{(a)}, \quad \mathbf{h}^{(b)} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_t^{(b)}, \tag{4}$$

where $\mathbf{h}^{(a)}, \mathbf{h}^{(b)} \in \mathbb{R}^{B\times D}$ are the utterance-level embeddings for the two views.

For each sample i in the batch, the two views $\mathbf{h}_i^{(a)}$ and $\mathbf{h}_i^{(b)}$ form a positive pair. All other embeddings in the batch—including both views of other samples—are treated as negative examples. The contrastive loss for a specific anchor view $k \in \{a, b\}$ of the i-th sample is defined as:

$$\ell_{i,k} = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{h}_i^k, \mathbf{h}_i^{k'})/\tau\right)}{\sum_{j=1}^{B}\sum_{u\in\{a,b\}}\mathbb{I}_{(j,u)\neq(i,k)}\exp\left(\mathrm{sim}(\mathbf{h}_i^k, \mathbf{h}_j^u)/\tau\right)}, \tag{5}$$

where k' denotes the other view of the same sample, $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity function, τ is a temperature hyperparameter, and $\mathbb{I}$ is an indicator function that excludes the anchor itself.

The total unsupervised contrastive loss is averaged over all $2B$ views in the batch:

$$\mathcal{L}_{\mathrm{UCL}} = \frac{1}{2B}\sum_{i=1}^{B}\left(\ell_{i,a} + \ell_{i,b}\right), \tag{6}$$

This loss function encourages the model to pull together representations of the same utterance under different dropout conditions while pushing apart representations from different utterances, thereby enhancing the discriminability of the learned speech features. Minimizing this objective essentially maximizes the mutual information between the perturbed views, forcing the encoder to filter out dropout-induced noise and retain invariant semantic structures, thereby learning a manifold robust to acoustic variations.

C. Objective Function

The overall training objective combines four loss components: the standard CTC loss $\mathcal{L}_{\mathrm{CTC}}$ [1], the attention-based decoder loss $\mathcal{L}_{\mathrm{att}}$ [4], the consistency regularization loss $\mathcal{L}_{\mathrm{KL}}$ (Section III-A), and the unsupervised contrastive loss $\mathcal{L}_{\mathrm{UCL}}$ (Section III-B).

For $\mathcal{L}_{\mathrm{CTC}}$, given a target transcript $\mathbf{y}$, the CTC loss derived from a forward pass with dropout mask $\boldsymbol{\xi}^{(s)}$ is calculated as

$\mathcal{L}_{\text{CTC}}^{(s)} = -\log P(\mathbf{y}|\mathbf{X}; \boldsymbol{\xi}^{(s)})$. We average the CTC losses over the two stochastic passes:

$$\mathcal{L}_{\text{CTC}} = \frac{1}{2}\left(\mathcal{L}_{\text{CTC}}^{(a)} + \mathcal{L}_{\text{CTC}}^{(b)}\right). \quad (7)$$

The final joint optimization objective is formulated as:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{att}} + \beta \cdot \mathcal{L}_{\text{KL}} + \gamma \cdot \mathcal{L}_{\text{UCL}}, \quad (8)$$

where $\mathcal{L}_{\text{att}}$ represents the standard cross-entropy loss, and λ balances the CTC and attention objectives (typically set to 0.3 [22]). The hyperparameters β and γ control the weights of consistency regularization and contrastive learning, respectively. This joint optimization enhances both alignment accuracy and the representation capability of the ASR system.

IV. Experiments

A. Experimental Settings

Datasets and Baselines. We evaluate RUCL on two standard benchmarks: the 178-hour Mandarin corpus AISHELL-1 [23] and the 960-hour English corpus LibriSpeech [24]. We compare with strong baselines including TDNN-LFMMI [23], SA-T [25], LAS [26], Transformer [27], Conformer [28], Wav2Vec 2.0 [8], HuBERT [9], TASA [29], and Diversity & Independence Regularization [7].

Implementation Details. Implemented in WeNet [22], we use 80-dimensional log-mel filterbanks with SpecAugment [16] and speed perturbation. The model has a 12-layer Conformer encoder ($d_{model} = 256, d_{ffn} = 2048$) and a 6-layer Transformer decoder. We use 4,231 characters for AISHELL-1 and 5,000 BPE tokens [30] for LibriSpeech. Models are trained on NVIDIA RTX 5090 GPUs with Adam optimizer and warmup. Key hyperparameters: CTC weight $\lambda = 0.3$, consistency weight $\beta = 0.4$, contrastive weight $\gamma = 0.3$. For inference, we use joint CTC/Attention decoding with beam size 10. CER and WER are used as evaluation metrics for AISHELL-1 and LibriSpeech, respectively.

B. Comparative experiment

Results on AISHELL-1. As shown in Table I, The vanilla Conformer achieved CERs of 5.05% (dev) and 5.78% (test). Its enhanced variants, R-TASA-Conformer and D-TASA-Conformer, showed better performance: their dev/test CERs were 4.96%/5.45% and 4.84%/5.21%, respectively. Our RUCL framework outperformed all these Conformer-based baselines. Compared with the strongest baseline D-TASA-Conformer, RUCL achieved an 11.13% relative CER.

Results on LibriSpeech. Table II shows comparative results on LibriSpeech. On Test-Clean, RUCL achieves a WER of 4.44%, outperforming Vanilla Conformer (5.33%) and D-TASA-Transformer (5.66%), showing effective representation refinement. On Test-Other, RUCL obtains 9.96% WER, significantly better than Wav2Vec 2.0 (13.3%). Unlike Ko et al. [7] (10.21%) with pre-trained backbone, RUCL is trained from scratch with a lightweight encoder and achieves lower error rates on noisy data. This proves our hierarchical constraints learn robust representations without large-scale pre-training.

TABLE I: Comparative experiments of different methods on the AISHELL-1 dataset.The LM indicates whether an additional language model is used during decoding.

Model	LM	Dev CER (%)	Test CER (%)
TDNN-LFMMI [23]	✓	6.44	7.62
SA-T [25]	×	8.30	9.30
LAS [26]	✓	–	8.71
ESPnet (Transformer) [27]	✓	6.00	6.70
Vanilla Transformer [31]	×	6.10	6.78
R-TASA-Transformer [29]	×	5.69	6.21
D-TASA-Transformer [29]	×	5.57	6.06
Vanilla Conformer [22]	×	5.05	5.78
R-TASA-Conformer [29]	×	4.96	5.45
D-TASA-Conformer [29]	×	4.84	5.21
RUCL (Ours)	×	**4.19**	**4.63**

TABLE II: Comparison on LibriSpeech without Language Model. The proposed RUCL employs a lightweight encoder ($d = 256$) trained from scratch, contrasting with the larger pre-trained architectures of SSL baselines ($d = 768$).

Model	Encoder Dim (d)	Labeled Data	Dev		Test	
			Clean	Other	Clean	Other
Self-Supervised Models						
Wav2Vec 2.0 [8]	768	100h	6.1	13.5	6.1	13.3
HuBERT [9]	768	100h	5.5	13.0	6.3	13.2
Ko et al. (CTC) [7]	768	960h	4.39	10.33	**4.34**	10.21
Supervised Models						
Vanilla Conformer	256	960h	4.81	10.80	5.33	10.86
R-TASA-Transformer [29]	256	960h	5.60	13.95	5.70	13.79
D-TASA-Transformer [29]	256	960h	5.58	13.58	5.66	13.59
Proposed						
RUCL (Ours)	256	960h	**4.08**	**9.68**	4.44	**9.96**

C. Noise Robustness Experiment

To evaluate robustness, we added Gaussian noise ($\sigma \in [0.1, 0.5]$) to AISHELL-1 test FBank features. As shown in Table III, performance degrades with noise for both models, while RUCL consistently outperforms Vanilla Conformer across all levels. Notably, at strong noise ($\sigma = 0.5$), RUCL keeps CER at 4.74%, better than the baseline at milder $\sigma = 0.3$. This excellent noise immunity confirms our hierarchical constraints act as internal noise injection, improving tolerance to external acoustic perturbations.

TABLE III: Comparison of CER (%) on AISHELL-1 test set under different Gaussian noise levels (σ).

Model	Noise Level (σ)				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
Vanilla Conformer	4.96	4.95	5.00	5.06	5.18
RUCL (Ours)	**4.57**	**4.57**	**4.63**	**4.65**	**4.74**

TABLE IV: Ablation study on the LibriSpeech dataset (WER, %).

Model	WER(%)			
	Dev Clean	Dev Other	Test Clean	Test Other
Conformer($\mathcal{L}_{\text{CTC}}$+$\mathcal{L}_{\text{att}}$) [22]	4.81	10.80	5.33	10.86
+$\mathcal{L}_{\text{KL}}$	4.51	10.32	4.87	10.37
+$\mathcal{L}_{\text{UCL}}$	4.32	10.46	4.73	10.30
+$\mathcal{L}_{\text{KL}}$+$\mathcal{L}_{\text{UCL}}$	**4.08**	**9.68**	**4.44**	**9.96**

D. Ablation Studies

Ablation studies on LibriSpeech (Table IV) verify the effectiveness of each component. Adding consistency regularization ($\mathcal{L}_{\text{KL}}$) or unsupervised contrastive learning ($\mathcal{L}_{\text{UCL}}$) alone reduces Test-Other WER from 10.86% to 10.37% and 10.30%. The joint optimization in RUCL achieves the best performance, with WERs reduced to 4.44% (Test-Clean) and 9.96% (Test-Other). This demonstrates that hierarchical fusion effectively enhances feature discriminability and prediction stability.

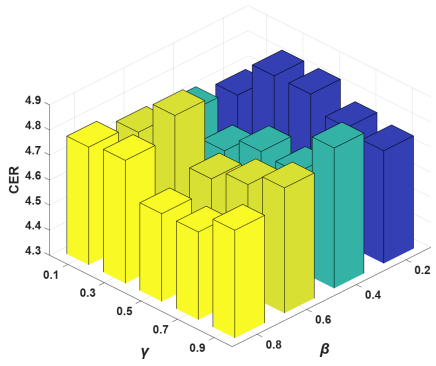
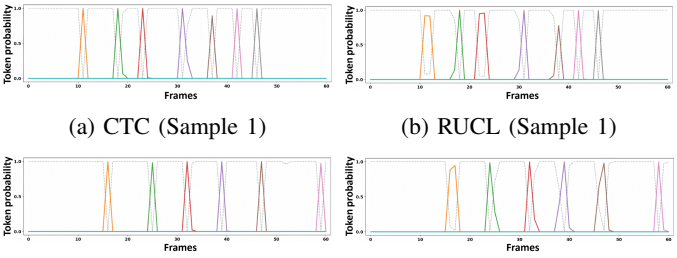


Fig. 2: CER (%) on AISHELL-1 test set under different hyperparameter combinations β, γ.

E. Hyperparameter Analysis

We performed a grid search on hyperparameters β and γ to evaluate their effects. As shown in Fig. 2, the model achieves optimal CER of 4.63% with $\beta = 0.4$ and $\gamma = 0.3$. Results show a stable region for $\beta \in [0.4, 0.6]$ and $\gamma \in [0.3, 0.5]$. Deviating causes performance drop: excessive β leads to over-regularization, while large γ overshadows the CTC objective. This proves balanced weights are essential to maximize the synergistic gains of RUCL.



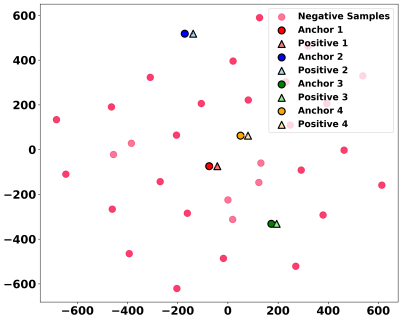
(a) CTC (Sample 1) (b) RUCL (Sample 1)

(c) CTC (Sample 2) (d) RUCL (Sample 2)

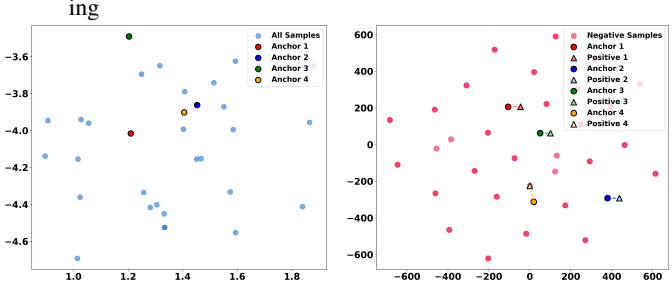
Fig. 3: Visualization of frame-level posterior probabilities comparison between Baseline CTC and RUCL. The solid colored lines represent non-blank tokens, while the gray dashed line represents the blank token.

F. Peak Suppression Analysis

We visualized frame-level posterior probabilities in Fig. 3. Baseline CTC (Fig. 3(a)(c)) shows sharp spikes on single frames, while RUCL (Fig. 3(b)(d)) yields much smoother distributions with wider temporal coverage. This suppression comes from consistency regularization: minimizing dropout divergence penalizes narrow overconfident predictions. The model learns temporal-shift robust representations. Such smoothness regularizes the model and reduces overfitting to training alignments. Combined with UCL discriminability, RUCL balances recognition precision and generalization, showing strong noise robustness.



(a) Regularization and unsupervised contrastive learning



(b) Vanilla Conformer (c) Consistency regularization

Fig. 4: t-SNE visualization of features learned by different models.

G. Visualization Analysis

To explore the reasons for RUCL's robustness, we visualized latent representations via t-SNE (Fig. 4). Vanilla Conformer (Fig. 4(b)) shows dispersed embeddings with blurry boundaries, indicating standard CTC training fails to learn compact representations and is sensitive to noise. Consistency regularization (Fig. 4(c)) brings some aggregation but weak separation between utterances. In contrast, RUCL (Fig. 4(a)) achieves high intra-sample compactness and inter-sample separability. By unsupervised contrastive learning, RUCL makes the encoder learn invariant semantic features. This clear structure avoids feature drift under interference, verifying that hierarchical constraints effectively refine feature space and ensure strong robustness and accuracy.

V. Conclusion

We propose RUCL, a hierarchical framework that unifies feature- and output-level consistency via intrinsic dropout. It effectively mitigates prediction inconsistency and weak feature

discriminability in ASR. Using dropout to build multi-view constraints, RUCL strengthens representation robustness and stabilizes outputs without extra parameters or latency. Experiments on AISHELL-1 and LibriSpeech show its superior performance over strong baselines. Visualization verifies that RUCL suppresses peaky CTC distributions, learns compact clusters, and yields strong noise immunity.

VI. ACKNOWLEDGMENTS

This work was supported by the National natural science foundation of China (No. 62571355).

REFERENCES

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[2] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[3] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[4] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[5] L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu *et al.*, "R-drop: Regularized dropout for neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 890–10 905, 2021.

[6] Z. Yao, W. Kang, X. Yang, F. Kuang, L. Guo, H. Zhu, Z. Jin, Z. Li, L. Lin, and D. Povey, "CR-CTC: Consistency regularization on CTC for improved speech recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: https://arxiv.org/pdf/2410.05101v4.pdf

[7] Y.-E. Ko, M.-H. Lee, D.-H. Kim, and J.-H. Chang, "Improving generalization of end-to-end ASR through diversity and independence regularization," in *Proceedings of the 26th Annual Conference of the International Speech Communication Association (Interspeech)*, Rotterdam, The Netherlands, 2025, pp. 3578–3582. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2025/ko25_interspeech.html

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[10] Z. Zhu, P. Zhou, D. Wang, L. Cheng, and J. Zhu, "Online multi-relational clustering with dominant view mining," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 34, 2026, pp. 29 232–29 240.

[11] Z. Zhu, P. Zhou, W. Du, S. Min, and J. Zhu, "Unsupervised semantic discovery via global and local semantic alignment in multimodal clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 41, 2026, pp. 35 248–35 256.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

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

[14] D. Jiang, W. Li, M. Cao, W. Zou, and X. Li, "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," *arXiv preprint arXiv:2010.13991*, 2020.

[15] Y. Gao, J. Feng, T. Wang, C. Deng, and S. Zhang, "A ctc triggered siamese network with spatial-temporal dropout for speech recognition," *arXiv preprint arXiv:2206.08031*, 2022.

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[17] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *The Twelfth International Conference on Learning Representations*, 2024.

[18] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 4280–4284.

[19] R. Huang, X. Zhang, Z. Ni, L. Sun, M. Hira, J. Hwang, V. Manohar, V. Pratap, M. Wiesner, S. Watanabe *et al.*, "Less peaky and more accurate ctc forced alignment by label priors," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 831–11 835.

[20] H. Ding, K. Chen, and Q. Huo, "Improving knowledge distillation of ctc-trained acoustic models with alignment-consistent ensemble and target delay," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 2561–2571, 2020.

[21] R. Takashima, L. Sheng, and H. Kawai, "Investigation of sequence-level knowledge distillation methods for ctc acoustic models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6156–6160.

[22] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit," in *Proc. Interspeech*, 2021, pp. 4054–4058.

[23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, 2017, pp. 1–5.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[25] Z. Tian, J. Yi, J. Tao, Y. Bai, and Z. Wen, "Self-Attention Transducers for End-to-End Speech Recognition," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2019, pp. 4395–4399.

[26] C. Shan, C. Weng, G. Wang, D. Su, M. Luo *et al.*, "Component Fusion: Learning Replaceable Language Model Component for End-to-End Speech Recognition System," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5631–5635.

[27] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[28] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[29] T.-H. Zhang, X. Qian, F. Chen, and X.-C. Yin, "Transmitted and Aggregated Self-Attention for Automatic Speech Recognition," in *Proceedings of the 25th Annual Conference of the International Speech Communication Association (Interspeech)*. Kos, Greece: ISCA, 2024, pp. 227–231. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2024/zhang24q_interspeech.html

[30] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa9-Abstract.html