SimLabel: Similarity-Weighted Iterative Framework for Multi-annotator Learning with Missing Annotations

Liyun Zhang
D3 Center, The University of Osaka
Japan

Zheng Lian
Institute of automation, Chinese academy of science
China

Hong Liu
Xiamen University
China

Takanori Takebe
Cincinnati Children's Hospital Medical Center
Japan

Yuta Nakashima
SANKEN, The University of Osaka
Japan

ABSTRACT

Multi-annotator learning (MAL) aims to model annotator-specific labeling patterns. However, existing methods face a critical challenge: they simply skip updating annotator-specific model parameters when encountering missing labels—a common scenario in real-world crowdsourced datasets where each annotator labels only small subsets of samples. This leads to inefficient data utilization and overfitting risks. To this end, we propose a novel similarity-weighted semi-supervised learning framework (SimLabel) that leverages inter-annotator similarities to generate weighted soft labels for missing annotations, enabling the utilization of unannotated samples rather than skipping them entirely. We further introduce a confidence-based iterative refinement mechanism that combines maximum probability with entropy-based uncertainty to prioritize predicted high-quality pseudo-labels to impute missing labels, jointly enhancing similarity estimation and model performance over time. For evaluation, we contribute a new multimodal multi-annotator dataset, AMER2, with high and more variable missing rates, reflecting real-world annotation sparsity and enabling evaluation across different sparsity levels.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; **Machine learning**; **Multi-task learning**.

KEYWORDS

Multi-annotator Learning, Missing Labels, Soft Label, Annotator Similarity, Semi-supervision

1 INTRODUCTION

Multi-annotator learning (MAL) has recently emerged as a research hotspot due to its relevance in subjective or nondeterministic tasks, such as medical diagnosis [14], visual perception [40], etc.. MAL aims to model annotator-specific labeling patterns [11, 34].

However, existing MAL methods face a critical challenge: they simply skip updating annotator-specific model parameters during training when encountering missing labels—a common scenario in real-world crowdsourced datasets, where each annotator labels only a small and often non-overlapping subset of samples to improve annotation efficiency [16]. This leads to low data utilization and increased risk of overfitting, as the annotator model is trained on limited annotations due to extensive missing labels.

To address this limitation, we propose a novel similarity-weighted semi-supervised learning framework (SimLabel), which estimates pairwise inter-annotator similarities and leverages them to generate weighted soft labels for missing annotations. It should be clarified that our goal is not to "fix" the inherent missing label characteristic of crowdsourced data, but rather to enable more effective data utilization. We aim to ensure that annotator-specific model parameters can still be updated when labels are missing, rather than

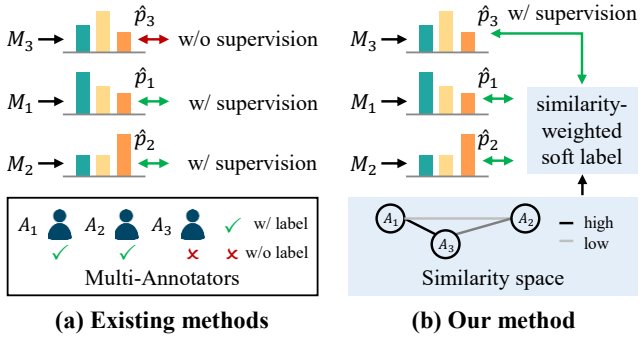

Figure 1: The sample is labeled by annotators A_1 to A_3, with A_3's label missing. (a) In existing methods, the predicted label distribution $\hat{p}_3$ from A_3's model M_3 lacks supervision due to the missing label, resulting in skipped parameter updates for M_3 on this sample. (b) In contrast, our method leverages labeling pattern similarities among A_1 to A_3, estimated from the dataset, to generate a soft label $\bar{p}_3$ that approximates A_3's true label. This is achieved via similarity-weighted aggregation of predictions $\hat{p}_1$ and $\hat{p}_2$, enabling semi-supervised updates of M_3 despite label missing.

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

simply skipped, thereby improving model performance through enhanced supervision.

Specifically, consider the example in Figure 1 with three annotators (A_1, A_2, and A_3), where A_3's annotation is missing for a given sample. As shown in Figure 1(a), existing approaches train separate models (M_1, M_2, and M_3) for each annotator. However, when A_3's label is absent, existing methods simply skip updating M_3's parameters entirely due to a lack of supervision, wasting valuable training opportunities and leading to inefficient data utilization and potential overfitting, as repeated application of this practice may result in annotator-specific models being trained on a small dataset with numerous missing labels.

In contrast, our method proposed in Figure 1(b) leverages Cohen's kappa coefficient [27] to calculate pairwise inter-annotator similarities. When A_3's label is missing, we weight the predicted distributions from other annotators (A_1 and A_2) based on their similarities to A_3, generating a similarity-weighted soft label $\tilde{p}_3$ to supervise model M_3's training. This semi-supervised approach enables continuous parameter updates rather than skipping missing annotations entirely, thereby improving data utilization efficiency and reducing overfitting risks.

Meanwhile, we introduce a confidence assessment mechanism that combines maximum probability values and entropy-based uncertainty metrics to evaluate the similarity-weighted soft labels. High-confidence predictions exceeding a predetermined threshold represent high reliability of the generated pseudo-labels, which are used to impute original missing labels in the dataset to recalculate the inter-annotator similarity matrix. This establishes a self-reinforcing cycle that jointly enhances similarity estimation and model performance over time.

To facilitate evaluation, we contribute a new multimodal multi-annotator dataset for video emotion recognition, AMER2, with 10 annotators, high and more variable missing rates across annotators (ranging from 75.9% to 91.3%). AMER2 better reflects real-world sparse annotation scenarios and enables evaluation under varying levels of label sparsity. Our contributions are as follows:

- **We propose a novel similarity-weighted semi-supervised learning framework** that addresses the missing label challenge in multi-annotator learning. It leverages inter-annotator similarities to generate weighted soft labels, enabling annotator-specific model updates when annotations are missing rather than skipping them entirely. This improves data utilization and reduces overfitting risk, enhancing model performance.
- **We introduce a confidence-based iterative refinement mechanism** that combines maximum probability with entropy-based uncertainty to dynamically prioritize predicted high-quality pseudo-labels to impute missing labels, jointly enhancing similarity estimation and model performance over time.
- **We contribute a new multimodal multi-annotator dataset, AMER2**, with high and more variable missing rates across 10 annotators (ranging from 75.9% to 91.3%), which better reflects real-world sparse annotation scenarios and enables evaluation under varying levels of label sparsity.

2 RELATED WORK

2.1 Traditional Multi-annotator Learning

Traditional multi-annotator learning aims to estimate a single consensus or ground-truth label from multiple annotators' labels. These include early probabilistic models [8], EM algorithms [17, 30], Gaussian models [19], CNN models [1], and biased estimation [29]. Tanno et al. [23] proposed modeling annotator confusion matrices as learnable parameters in neural networks. Cao et al. [4] introduced max-MIG to learn from multiple annotators. NEAL [5] employs neural expectation-maximization to jointly learn annotator expertise and true labels. Later methods used probabilistic frameworks to aggregate multiple annotations into a consensus or ground-truth label by confusion matrix [24], agreement distribution [28], and Gaussian distributions [14]. This aggregation paradigm often treats annotator disagreements as noise to be averaged away rather than valuable information [12, 31]. The underlying computer vision and machine learning techniques used in multi-annotator learning have broader applications across various domains [21, 35, 37**?** –39], though annotator disagreements in multi-annotator scenarios reflect genuine perspective differences rather than noise.

2.2 Multi-annotator Labeling Pattern Modeling

Some studies have also attempted to model individual annotator patterns and provide explanations: D-LEMA [15] trains annotator models on non-contradictory subsets with spatial weights for noise handling; MaDL [11] jointly optimizes ground truth classifiers and annotator models through weighted embeddings; TAX [6] associates convolutional kernels with prototype libraries for pixel-level annotation decisions; MAGI [41] leverages annotator explanations to address noisy annotations, and Schaekermann et al. [20] analyze factors contributing to disagreements. Particularly, QuMATL [34] models individual annotator labeling patterns via learnable queries with behavioral analysis, introducing a paradigm shift, i.e., views each annotator as having unique labeling patterns worth preserving rather than as noisy approximations of ground truth. However, missing labels are inherent characteristics of crowdsourced data; the annotator-specific model parameters will skip update in case of missing labels, thereby influencing the individual annotator modeling.

2.3 Multi-annotator Learning with Missing Labels

To the best of our knowledge, the multi-annotator learning with missing labels task problem has not yet been investigated. Several similar works are as follows: Yan et al. [32] modeled annotator expertise for complete annotation scenarios without addressing missing annotation challenges. Davani et al. [7] preserved disagreements beyond majority voting but lacked a framework for missing annotator labels. Li et al. [13] incorporated instance features but assumed available annotations, while Tanno et al. [25] addressed noisy labels through confusion estimation without handling missing data. Guan et al. [9] demonstrated individual labeler modeling benefits but did not consider missing label scenarios. Shah et al. [22] explored crowdsourcing self-correction mechanisms without specific missing label strategies. Rodrigues et al. [18] proposed deep

Table 1: Label statistics and missing rates of the AMER2 dataset compared to the AMER dataset. For each annotator, the number of labeled samples, the corresponding missing rate (%), as well as the average data, and the total number of samples are reported. AMER contains 13 annotators while AMER2 contains 10 annotators, denoted as A_k.

Number of samples	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	Average	Total
AMER	1096	1031	1022	1036	1012	970	1064	1049	1060	1062	5187	5197	5202	1999.1	5207
AMER2	545	538	201	557	493	545	346	544	542	545	-	-	-	485.6	2311
Missing rate (%)	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	Average	-
AMER	79.0	80.2	80.4	80.1	80.6	81.4	79.6	79.9	79.6	79.6	0.4	0.2	0.1	69.6	-
AMER2	76.4	76.7	91.3	75.9	78.7	76.4	85.0	76.5	76.5	76.4	-	-	-	79.0	-

learning from crowds assuming complete crowd annotations, and Albarqouni et al. [2] introduced AggNet for medical imaging but focused on aggregating available annotations rather than handling missing ones.

Existing multi-annotator learning methods directly skip annotator-specific model parameter updates in case of missing labels, which not only causes inefficient data utilization but may also lead to overfitting of annotator models on small datasets. Our work fills this critical gap by proposing a framework that leverages the similarity relationships between annotators to achieve annotator-specific model parameter updates when labels are missing rather than simply skipping. By introducing this similarity-based soft constraint for cases of missing labels, our approach avoids the low data utilization efficiency and potential overfitting risks caused by skipping annotator-specific model parameter updates due to a lack of supervision when labels are missing.

3 DATASET CONSTRUCTION

This paper introduces a new video emotion recognition dataset, AMER2, which is an extended version of the AMER dataset [34]. The AMER dataset contains 5,207 video samples and provides rich per-annotator labels to meet the requirements of multi-annotator tendency learning [34].

Unlike AMER, AMER2 provides an additional 2,311 samples and sparse per-annotator labels, with the intention of validating the effectiveness of our proposed method under more challenging missing conditions. In AMER2, most samples focus on single-person videos with relatively complete speech content, sourced from movies and TV series.

During the annotation process, we utilize the Label Studio toolkit [26] and hire multiple annotators who are masters or PhD students in our labs. To ensure annotation quality, these annotators first undergo preliminary exams. In these exams, we provide 10 samples and ask the annotators to select the most likely label from 8 candidate labels: *worry*, *happiness*, *neutral*, *anger*, *surprise*, *sadness*, *other*, and *unknown*. These samples were previously annotated by five experts and have obtained five-agreement labels. Annotators who fail to pass the preliminary exam are removed from the annotation pool. After that, we retained 10 annotators, and each annotator completed the task in approximately two weeks, with scheduled breaks to maintain annotation quality. Finally, each annotator provided approximately 201 to 557 labels.

Table 1 provides statistics for AMER and AMER2. From this table, we observe that AMER2 has an overall average missing rate of 79.0%, with one annotator's missing labels reaching an extreme of 91.30%, which is higher than AMER's overall average missing rate of 69.6%. Therefore, AMER2 increases the proportion of missing labels to better mimic real-world scenarios with sparse per-annotator labels. In this paper, we conduct experiments on both datasets, aiming to validate the effectiveness of our method under variable missing rates.

4 METHODOLOGY

SimLabel contains two components: the similarity-weighted framework and confidence-based iterative refinement. Similarity-weighted framework generates soft labels for missing annotations through inter-annotator similarity weights, providing semi-supervised constraints for annotator-specific models (Figure 2). Confidence-based iterative refinement dynamically updates the similarity matrix by evaluating confidence scores of generated soft labels, creating a self-reinforcing learning cycle (Figure 3).

4.1 Similarity-weighted Framework

We propose a novel approach to address the issue of multi-annotator learning with missing labels. As shown in Figure 2, our dataset consists of pairs of a video x and a set $\mathcal{Y} = \{y\}$ of labels $y \in \{0, 1\}^N$ in the one-hot representation, where N is the number of classes. $\mathcal{Y}$ contains labels by multiple annotators A_k ($k = 1, \ldots, K$, where K is the number of annotators), providing different annotations for the same video because of the subjectivity or nondeterministicity of the task. Often, not all annotators label all samples (thus $|\mathcal{Y}| \leq K$), resulting in missing labels—a common situtaion in real-world scenarios. In this example, for emotion assessment in the video, A_1 gives the label *'neutral'*, A_2 gives the label *'sad'*, while A_3 does not give the label, i.e, the label is missing.

The video x is processed by separate classification model M_k for each annotator A_k, designed to learn individual labeling patterns. The choice of these classification models are arbitrary: They can use Gaussian distribution fitting (PADL [14]), confusion matrix (MaDL [11]), or query-based architecture (QuMATL [34]), etc. to model individual annotators (and their relationship). Given an input data x, each model produces label distribution $\hat{p}_k(l|x)$ for A_k and class l ($l = 1, \ldots, N$).

When the annotation from A_k is available for pair $(x, \mathcal{Y}$, we update the corresponding model M_k using $y_k \in \mathcal{Y}$ through supervised

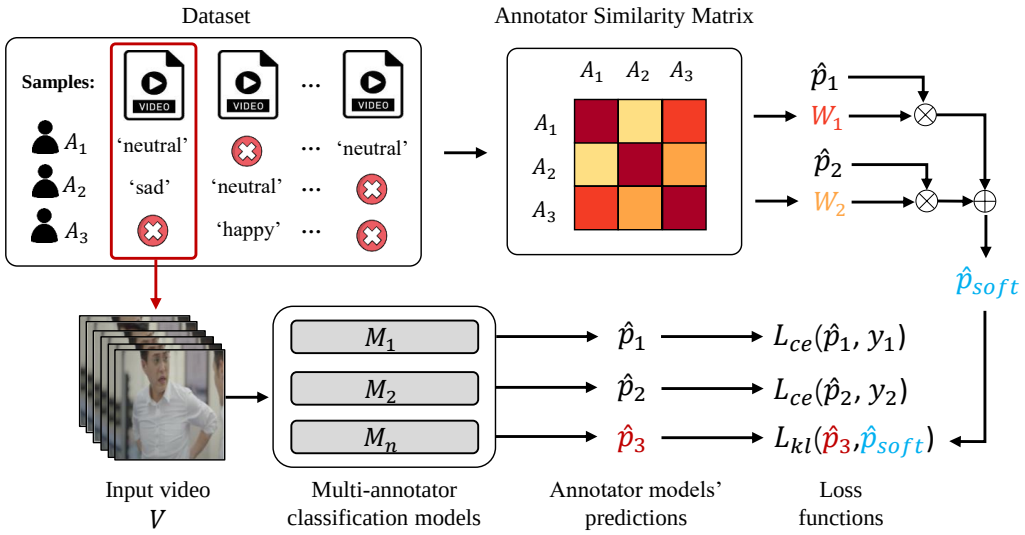

Figure 2: The main framework of SimLabel. For datasets with missing labels, annotator similarity is calculated via the Cohen's kappa coefficient (darker colors indicate higher similarity). Using video sample V with missing label from A_n: Multiple annotator-specific models process V to produce label distributions. Labeled predictions ($\hat{p}_1$, $\hat{p}_2$) are supervised through cross-entropy loss with ground truth labels (y_1, y_2). Unlabeled predictions ($\hat{p}_n$) is constrained via KL divergence loss with a soft label ($\hat{p}_{soft}$), generated as a similarity-weighted combination of $\hat{p}_1$ and $\hat{p}_2$ using weights W_1 and W_2 derived from annotator similarities to A_n.

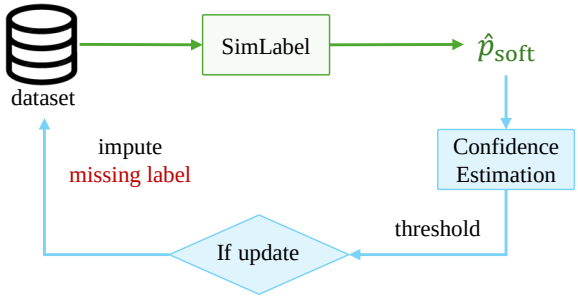

Figure 3: Confidence-based Iterative Refinement. Confidence is calculated for the soft label ($\bar{p}_k$) generated for missing labels. When confidence exceeds a predetermined threshold, the predicted high-quality pseudo-label imputes the original missing label in the dataset, iteratively updating the inter-annotator similarity matrix and utilized for soft label ($\bar{p}_k$) generation in subsequent iterations.

learning by computing cross-entropy loss:

$$\mathcal{L}_{\text{ce}}(\hat{p}_k, y_k) = -\sum_{l=1}^{N} y_{kl} \log \hat{p}_k(l|x). \tag{1}$$

When annotation from A_k is unavailable, we update model M_k in a semi-supervised manner by computing Kullback-Leibler (KL) divergence loss with generating a soft label $\bar{p}_k \in [0,1]^N$:

$$\mathcal{L}_{\text{kl}}(\hat{p}_k, \bar{p}_k) = D_{\text{KL}}(\hat{p}_k \| \bar{p}_k). \tag{2}$$

The soft label $\bar{p}_k$ is key to our method. Based on our assumption that inter-annotator correlations derived from labels in a multi-annotator dataset can give some ideas on missing labels, inter-annotator similarities across the entire dataset to indirectly update models with unannotated samples through semi-supervised learning. We calculate the similarity matrix between annotators using the Cohen's kappa coefficient [27] from the original dataset. Figure 2 illustrates the matrix, where darker colors indicate greater similarity between annotators. A_1 has higher similarity to A_3 compared to A_2; therefore, when A_3's label is missing, A_1's label y_1 may be more informative to predict y_3 compared A_2's. We define the similarity weights of A_1 relative to A_3 and A_2 relative to A_3 as $w_{1,3}$ and $w_{2,3}$, respectively. The soft label $\bar{p}_3$ is thus generated by the weighted sum of label distribution predictions $\hat{p}_1$ and $\hat{p}_2$ with their corresponding similarity weights $w_{1,3}$ and $w_{2,3}$. In general, we generate the soft label for A_k by:

$$\bar{p}_k = \sum w_{k',k} \hat{p}_{k'}, \tag{3}$$

where the summation is computed over all $k' \neq k$. Here, k' refers to the number of those annotators who have labels; k refers to the annotator who has the missing label.

4.2 Confidence-based Iterative Refinement

Building upon the similarity-weighted framework, we introduce a confidence assessment mechanism for the similarity-weighted soft labels. As shown in Figure 3, if the confidence score exceeds a predetermined threshold, indicating high reliability of the generated soft label, the predicted label will impute the corresponding missing labels of the original dataset to recalculate the inter-annotator

Algorithm 1 The Confidence-based Iterative Refinement for Dynamic Inter-Annotator Similarity Relationship

Require: Dataset $\mathcal{D}$ with missing labels, confidence threshold T
1: Initialize similarity matrix SM using Cohen's kappa coefficient on available labels
2: Initialize annotator models $\{M_1, M_2, \ldots, M_K\}$
3: **for** each training epoch **do**
4: **for** each sample with missing labels **do**
5: Generate similarity-weighted soft label: $\bar{p}_k = \sum w_{k',k} \hat{p}_{k'}$
6: Calculate confidence: $c = \max(\bar{p}_k) \times (1 - \frac{H[\bar{p}_k]}{H_{\max}})$
7: **if** $c \geq T$ **then**
8: Extract predicted label: $y_{pred} = \arg\max(\bar{p}_k)$
9: Impute y_{pred} into dataset for corresponding missing annotation
10: Recalculate annotator similarity matrix SM using updated dataset
11: **end if**
12: Update annotator models using supervised and semi-supervised losses
13: **end for**
14: **end for**
Ensure: Refined similarity matrix SM, imputed dataset, trained annotator models

similarity matrix. Through this mechanism, we establish a self-reinforcing cycle that continuously refines the inter-annotator similarity relationships and the individual annotator models' ability, leading to progressively more accurate predictions throughout the training process. This mechanism integrates both maximum probability values and entropy-based uncertainty metrics to provide comprehensive confidence estimates and identify highly reliable soft labels.

Confidence Calculation. As shown in Algorithm 1, for the similarity-weighted soft labels $\bar{p}_k$ generated for a missing annotation of A_k, we perform confidence calculations to enhance our semi-supervised method. Our confidence combines maximum probability with normalized entropy to provide a comprehensive assessment of prediction reliability:

$$c = \max(\bar{p}_k) \times (1 - \frac{H[\bar{p}_k]}{H_{\max}}), \tag{4}$$

where $\max(\bar{p}_k)$ is the maximum value in $\bar{p}_k$, $H[\bar{p}_k]$ is the entropy of $\bar{p}_k$, and $H_{\max} = \log N$. The right side of the multiplication is to normalize c into $[0, 1]$.

This formulation requires predictions to have both high maximum probability and low normalized entropy to achieve high confidence scores. This provides a comprehensive assessment of prediction reliability by balancing two key factors: (1) The maximum probability term $\max(\bar{p}_k)$ captures the model's confidence in the most likely class. (2) The normalized entropy term $H_{norm}(\bar{p}_k) = (1 - \frac{H[\bar{p}_k]}{H_{\max}})$ measures the uncertainty across the entire distribution, with lower values indicating more concentrated (certain) predictions.

Dynamic Refinement Process. As shown in Algorithm 1, when the calculated confidence exceeds a predetermined threshold T (Algorithm 1, line 7), indicating that the generated soft label has high reliability, the predicted label y_{pred} is extracted and incorporated into the location of missing label in the original dataset (lines 8–9). The annotator similarity matrix SM is then recalculated using the updated dataset with newly imputed labels (line 10).

This process facilitates more accurate establishment of similarity relationships between annotators in cases of missing labels. As training progresses and missing labels meeting confidence criteria are incorporated, a virtuous cycle emerges, continuously refining the similarity relationships between annotators. The dynamic refinement allows each annotator model to more accurately capture its specific annotation patterns, even when starting from datasets with significant numbers of missing annotations.

The effectiveness of this approach lies in its ability to leverage high-confidence predictions to bootstrap the learning process, creating a self-improving system where each iteration potentially enhances the quality of both the similarity matrix and the generated soft labels for remaining missing annotations.

5 EXPERIMENT

We conduct extensive experiments comparing SimLabel (with and without confidence-based iterative refinement) against existing approaches that simply skip annotator-specific model parameter updates in case of missing labels. To verify effectiveness across different architectures, we evaluate on three annotator modeling frameworks: Gaussian distribution fitting (PADL [14]), confusion matrix (MaDL [11]), and query-based modeling (QuMATL [34]). Experiments are conducted on AMER2 and AMER containing real missing labels, and the STREET dataset with simulated missing labels at various missing ratios, using Accuracy and Difference of Inter-annotator Consistency (DIC) [34] as evaluation metrics.

Note that due to space limitations, we discuss additional key issues with experimental results, including training dynamics, threshold sensitivity to missing rates, and strategies for handling noisy labels and avoiding propagation errors, etc., in the **supplementary material**.

5.1 Implementation Details

We use Cohen's kappa coefficient [27] to calculate the inter-annotator similarity matrix. Image and video data are all resized to 224×224 and further normalized. For different annotator model architectures, we follow their original training and testing settings. These experiments are achieved on four NVIDIA V100 GPUs.

5.2 Evaluation Metrics

Accuracy is a standard metric to evaluate individual annotator modeling. DIC [34] quantifies how inter-annotator correlations differ between ground-truths and predictions, and we also use DIC to evaluate our approach's benefits from the perspective of inter-annotator consistency.

5.3 Datasets

For the dataset, we primarily utilize the newly constructed multimodal emotion recognition dataset AMER2, alongside the earlier

Table 2: Accuracy comparison on AMER2 dataset (10 annotators, A_k, $k = 1, \ldots, 10$) for annotator modeling performance with average (Avg). Methods compared: existing approach (Architecture - Skip); similarity-weighted framework (Architecture - Ours); similarity-weighted framework with confidence-based iterative refinement (Architecture - Ours + Confidence).

Methods	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	Avg
PADL - Skip	0.81	0.84	0.82	0.87	0.78	0.80	0.78	0.84	0.80	0.79	0.81
PADL - Ours	0.84	0.85	**0.84**	0.89	0.82	0.81	0.82	0.85	0.82	0.83	0.84
PADL - Ours + Confidence	**0.86**	**0.87**	**0.84**	**0.90**	**0.84**	**0.83**	**0.85**	**0.86**	**0.85**	**0.86**	**0.86**
MaDL - Skip	0.84	0.83	0.82	0.86	0.81	0.82	0.80	0.82	0.85	0.84	0.83
MaDL - Ours	0.87	0.84	0.85	0.87	0.83	0.85	0.82	0.86	0.86	0.85	0.85
MaDL - Ours + Confidence	**0.89**	**0.86**	**0.88**	**0.88**	**0.86**	**0.87**	**0.84**	**0.88**	**0.88**	**0.87**	**0.87**
QuMATL - Skip	0.86	0.83	0.87	0.84	0.86	0.87	0.88	0.83	0.85	0.86	0.86
QuMATL - Ours	**0.89**	0.85	0.90	0.86	0.89	0.88	0.91	**0.86**	0.87	0.89	0.88
QuMATL - Ours + Confidence	**0.89**	**0.87**	**0.92**	**0.88**	**0.91**	**0.90**	**0.93**	**0.86**	**0.90**	**0.91**	**0.90**

Table 3: Accuracy comparison on AMER dataset evaluating annotator modeling performance for 13 annotators.

Methods	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	Avg
PADL - Skip	0.89	0.90	0.88	0.93	0.87	0.91	0.86	0.94	0.89	0.88	0.47	0.54	0.35	0.79
PADL - Ours	0.91	0.92	0.90	**0.94**	0.90	0.92	0.89	0.95	0.91	0.91	0.55	0.61	0.45	0.83
PADL - Ours + Confidence	**0.92**	**0.93**	**0.91**	**0.94**	**0.91**	**0.93**	**0.90**	**0.96**	**0.92**	**0.93**	**0.59**	**0.65**	**0.50**	**0.85**
MaDL - Skip	0.93	0.91	0.90	0.89	0.90	0.88	0.90	0.89	0.87	0.92	0.50	0.53	0.37	0.80
MaDL - Ours	0.95	0.92	0.92	0.91	0.92	0.90	0.92	0.92	0.90	**0.94**	0.59	0.60	0.48	0.84
MaDL - Ours + Confidence	**0.96**	**0.93**	**0.93**	**0.92**	**0.93**	**0.91**	**0.93**	**0.93**	**0.91**	**0.94**	**0.64**	**0.65**	**0.53**	**0.86**
QuMATL - Skip	0.94	0.93	0.93	0.94	0.94	0.92	0.93	0.95	0.93	0.93	0.59	0.61	0.40	0.84
QuMATL - Ours	0.96	0.94	**0.95**	0.95	0.95	0.94	0.95	0.96	0.95	0.95	0.68	0.69	0.52	0.88
QuMATL - Ours + Confidence	**0.97**	**0.95**	**0.95**	**0.96**	**0.96**	**0.95**	**0.96**	**0.97**	**0.96**	**0.96**	**0.72**	**0.73**	**0.57**	**0.89**

Table 4: Randomly removing annotations at 40% missing ratios is to simulate sparser missing scenarios on STREET, AMER, and AMER2 datasets. -Ha, -He, -Sa, -Li, and -Or represent five perspectives: happiness, healthiness, safety, liveliness, and orderliness. The average modeling performance of whole annotators is evaluated by the accuracy metric.

Methods	STREET-Ha	STREET-He	STREET-Sa	STREET-Li	STREET-Or	AMER	AMER2
PADL - Skip	0.43	0.42	0.38	0.41	0.40	0.58	0.61
PADL - Ours	0.48	0.46	0.42	0.47	0.45	**0.64**	0.66
PADL - Ours + Confidence	**0.51**	**0.49**	**0.45**	**0.50**	**0.48**	**0.64**	**0.69**
MaDL - Skip	0.42	0.43	0.36	0.38	0.36	0.63	0.65
MaDL - Ours	**0.45**	0.46	0.38	0.41	0.39	0.66	0.69
MaDL - Ours + Confidence	**0.45**	**0.48**	**0.41**	**0.43**	**0.42**	**0.71**	**0.72**
QuMATL - Skip	0.52	0.51	0.46	0.49	0.48	0.66	0.68
QuMATL - Ours	0.56	0.55	0.50	0.53	0.52	0.70	0.71
QuMATL - Ours + Confidence	**0.60**	**0.58**	**0.54**	**0.57**	**0.56**	**0.74**	**0.75**

version AMER and the city impression assessment dataset STREET [34]. AMER2 and AMER naturally contain missing labels in real-world settings, while STREET is a complete real-world dataset. The AMER dataset's complexity in capturing temporal emotion dynamics aligns with recent advances in time-sensitive emotion recognition [33, 36]. For the STREET dataset, we randomly remove annotations at different missing ratios to simulate annotation absence. Similarly, we also apply this random removal procedure to AMER2 and AMER datasets to further increase missing rates and validate the effectiveness of our approach.

Table 5: DIC measures how well enhanced annotator modeling via missing label handling improves inter-annotator consistency toward ground truth, lower values indicate better gains. -S and -O represent Skip and Ours approaches.

Datasets	PADL-S	PADL-O	QuMATL-S	QuMATL-O
STREET-Ha	0.48	**0.44**	0.43	**0.38**
STREET-He	0.52	**0.45**	0.38	**0.34**
STREET-Sa	0.32	**0.28**	0.24	**0.20**
STREET-Li	0.43	**0.38**	0.27	**0.22**
STREET-Or	0.57	**0.53**	0.54	**0.49**
AMER	0.36	**0.32**	0.23	**0.19**
AMER2	0.34	**0.31**	0.22	**0.17**

5.4 Results Analysis

Table 2 and Table 3 present accuracy results for each annotator across different annotator model architectures based on the comparison between our proposed SimLabel (i.e., using only the similarity-weighted framework of the similarity-weighted soft label, defined as "- Ours", and using both the similarity-weighted framework and confidence-based iterative refinement, defined as "- Ours + Confidence") with existing approaches (i.e., directly skipping annotator-specific model parameter updates in case of missing labels, defined as "- Skip"). Table 4 presents average accuracy results for multi-annotators across different annotator model architectures at different missing labels.

On the AMER2 dataset (Table 2), our approach using the similarity-weighted framework ("- Ours") consistently outperforms existing approaches ("- Skip") which directly skip annotator-specific model parameter updates in case of missing labels, with average improvements of 3% for PADL, 2% for MaDL, and 2% for QuMATL. When incorporating confidence-based iterative refinement ("- Ours + Confidence") for our approach, we observe further enhancements of 2% across all different architectures.

Results on the AMER dataset (Table 3) show similar patterns of improvement. Our approach using the similarity-weighted framework ("- Ours") improves average accuracy by around 3% for PADL, MaDL, and QuMATL compared to existing approaches ("- Skip"). With confidence-based iterative refinement ("- Ours + Confidence") for our approach, these improvements further increase to around 2%. This consistent enhancement across different model architectures of multi-annotator learning validates our central hypothesis that leveraging inter-annotator similarity provides an effective framework for addressing missing label challenges in multi-annotator learning.

For the STREET dataset, we conducted experiments with artificially induced missing labels at different rates, we show 40% (Table 4) here, and more data are provided in the supplementary material. To evaluate robustness, we also apply this random removal procedure to AMER2 and AMER datasets to further increase missing rates and validate the effectiveness of our approach. Our approach delivers consistent improvements across all scenarios, which demonstrates that our method is particularly valuable in scenarios with severe label sparsity.

Table 6: Ablation studies on similarity matrix calculation, confidence threshold, and calculation choices on AMER2 dataset. Top: Similarity matrix calculation choice. Middle: Performance with different confidence thresholds. Bottom: Comparison of different confidence calculation methods.

Similarity matrix calculation	PADL	MaDL	QuMATL
Pearson correlation	0.82	0.84	0.86
Krippendorff's alpha	0.84	0.85	0.88
Cohen's kappa	**0.85**	**0.87**	**0.90**
Confidence Threshold	PADL	MaDL	QuMATL
$\tau = 0.5$	0.85	0.86	0.89
$\tau = 0.6$	**0.86**	**0.87**	**0.90**
$\tau = 0.7$	0.84	0.85	0.88
$\tau = 0.8$	0.82	0.83	0.86
Confidence Calculation	PADL	MaDL	QuMATL
$\max(\bar{p}_k)$	0.83	0.84	0.87
$(1 - \frac{H[\bar{p}_k]}{H_{\max}})$	0.84	0.85	0.88
$\max(\bar{p}_k) \times (1 - \frac{H[\bar{p}_k]}{H_{\max}})$	**0.86**	**0.87**	**0.90**

For the gain evaluation of inter-annotator consistency, the DIC scores in Table 5 also show that our similarity-weighted approach shows consistent gains compared to the skip way (i.e., skipping annotator-specific model parameter updates in case of missing labels) across different architectures on different datasets. Lower DIC values indicate that our approach better captures individual annotators' labeling patterns through enhanced annotator modeling via missing label handling, thereby improving inter-annotator consistency convergence toward ground truth. More detailed Table data is provided in the supplementary material.

These results consistently demonstrate that leveraging annotator similarity relationships through our soft label generation and confidence-based iterative refinement mechanism improves multi-annotator modeling performance, especially in realistic scenarios with missing annotations.

5.5 Ablation Study

To evaluate the design choices in our confidence-based iterative refinement mechanism, we conduct a detailed ablation study examining confidence threshold selection, its sensitivity to different missing rates, and the effectiveness of different confidence formulation methods. They are all performed on AMER2 dataset.

Similarity Matrix Calculation. We first need to clarify a key point: our confidence threshold does not "filter out erroneous pseudo-labels", but rather uses reliable (high-confidence) predictions to progressively refine the similarity matrix. Therefore, under the widely validated Cohen's kappa coefficient [27] and self-iterative framework, the model performance demonstrates robustness. Second, we conducted ablation experiments comparing Cohen's kappa with Pearson correlation coefficient [3] and Krippendorff's alpha coefficient [10] to evaluate the accuracy of the similarity matrix, as shown in Table 6 (top). Results on the AMER2 dataset show Cohen's kappa consistently outperforms alternatives through

chance agreement correction for categorical annotations. Pearson correlation coefficient fails to capture discrete characteristics, while Krippendorff's alpha shows instability in sparse scenarios. Even with these less-matched metrics, model performance degradation is minimal, validating similarity matrix robustness.

Confidence Threshold Selection. Table 6 (middle) shows the performance of our method with different confidence thresholds. The results indicate that a threshold of $\tau = 0.6$ achieves the best performance across all model architectures. Higher thresholds ($\tau = 0.8$) lead to performance degradation, likely because too few predictions meet the criteria for updating the similarity matrix. Lower thresholds ($\tau = 0.5$) also perform slightly worse than $\tau = 0.6$, possibly due to the inclusion of lower-quality predictions that introduce noise into the update process.

Confidence Formulation Comparison. Finally, we evaluate different confidence calculation methods (Table 6, bottom): (1) using only maximum probability $\max(\bar{p}_k)$, (2) using only normalized entropy complement $(1 - \frac{H[\bar{p}_k]}{H_{\max}})$, and (3) our proposed combined approach $\max(\bar{p}_k) \times (1 - \frac{H[\bar{p}_k]}{H_{\max}})$, where $\bar{p}_k$ represents the similarity-weighted soft label probability distribution generated for missing annotations. The results demonstrate that our combined method consistently outperforms single-metric approaches across all architectures, with an average performance improvement of 3% over $\max(\bar{p}_k)$ and 2% over entropy-only formulation. This validates our hypothesis that effective confidence assessment should consider both the strength of the dominant class prediction and the overall distribution shape.

6 CONCLUSION

We addressed the challenge of missing labels in multi-annotator learning through a similarity-weighted semi-supervised framework that leverages inter-annotator relationships instead of skipping annotator-specific model parameter updates in case of missing labels. SimLabel combines soft label generation with a confidence-based iterative refinement mechanism to dynamically refine inter-annotator similarity estimates. We also contribute a new dataset, AMER2, with high and variable missing rates to reflect real-world annotation sparsity and enable evaluation across different sparsity levels. Extensive experiments on different datasets and missing rates validated SimLabel's effectiveness to address the missing label challenge. In future work, we plan to extend this framework to handle dynamic annotator behaviors and explore more complex scenarios.

REFERENCES

[1] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. 2016. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging* 35, 5 (2016), 1313–1321.

[2] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. 2016. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging* 35, 5 (2016), 1313–1321.

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.

[4] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. 2019. Max-mig: an information theoretic approach for joint learning from crowds. *arXiv preprint arXiv:1905.13436* (2019).

[5] Junfan Chen, Richong Zhang, Jie Xu, Chunming Hu, and Yongyi Mao. 2023. A Neural Expectation-Maximization Framework for Noisy Multi-Label Text Classification. *IEEE Transactions on Knowledge and Data Engineering* 35, 11 (2023), 10992–11003. https://doi.org/10.1109/TKDE.2022.3223067

[6] Yuan-Chia Cheng, Zu-Yun Shiau, Fu-En Yang, and Yu-Chiang Frank Wang. 2023. TAX: Tendency-and-Assignment Explainer for Semantic Segmentation with Multi-Annotators. *arXiv preprint arXiv:2302.09561* (2023).

[7] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110.

[8] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[9] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[10] Kilem L Gwet. 2011. On the Krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October* 2, 2011 (2011), 2011.

[11] Marek Herde, Denis Huseljic, and Bernhard Sick. 2023. Multi-annotator Deep Learning: A Probabilistic Framework for Classification. *arXiv preprint arXiv:2304.02539* (2023).

[12] Uthman Jinadu, Jesse Annan, Shanshan Wen, and Yi Ding. 2023. Loss Modeling for Multi-Annotator Datasets. *arXiv preprint arXiv:2311.00619* (2023).

[13] Jingzheng Li, Hailong Sun, Jiyi Li, Zhijun Chen, Renshuai Tao, and Yufei Ge. 2021. Learning from multiple annotators by incorporating instance features. *arXiv preprint arXiv:2106.15146* (2021).

[14] Zehui Liao, Shishuai Hu, Yutong Xie, and Yong Xia. 2024. Modeling annotator preference and stochastic annotation error for medical image segmentation. *Medical Image Analysis* 92 (2024), 103028.

[15] Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. 2021. D-lema: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1837–1846.

[16] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419. https://doi.org/10.1017/S1930297500002205

[17] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *Journal of Machine Learning Research* 11, 43 (2010), 1297–1322. http://jmlr.org/papers/v11/raykar10a.html

[18] Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[19] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. Gaussian Process Classification and Active Learning with Multiple Annotators. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 433–441. https://proceedings.mlr.press/v32/rodrigues14.html

[20] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 76 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359178

[21] Xuanmeng Sha, Liyun Zhang, Tomohiro Mashita, and Yuki Uranishi. 2024. 3DFacePolicy: Speech-Driven 3D Facial Animation with Diffusion Policy. *arXiv preprint arXiv:2409.10848* (2024).

[22] Nihar Shah and Dengyong Zhou. 2016. No oops, you won't do it again: mechanisms for self-correction in crowdsourcing. In *International conference on machine learning*. PMLR, 1–10.

[23] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. 2019. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. 2019. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11244–11253.

[25] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. 2019. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11244–11253.

[26] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. https://github.com/HumanSignal/label-studio Open source software available from https://github.com/HumanSignal/label-studio.

[27] AnthonyJ. Viera and JoanneM. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine,Family Medicine* (May 2005).

[28] Chongyang Wang, Yuan Gao, Chenyou Fan, Junjie Hu, Tin Lum Lam, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2023. Learn2agree: Fitting with multiple annotators without objective ground truth. In *International Workshop on Trustworthy Machine Learning for Healthcare*. Springer, 147–162.

[29] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems* 23 (2010).

[30] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009).

[31] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. 2014. Learning from multiple annotators with varying expertise. *Machine learning* 95 (2014), 291–327.

[32] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. 2014. Learning from multiple annotators with varying expertise. *Machine learning* 95 (2014), 291–327.

[33] Liyun Zhang. 2024. MicroEmo: Time-Sensitive Multimodal Emotion Recognition with Micro-Expression Dynamics in Video Dialogues. *arXiv preprint arXiv:2407.16552* (2024).

[34] Liyun Zhang, Zheng Lian, Hong Liu, Takanori Takebe, and Yuta Nakashima. 2025. QuMATL: Query-based Multi-annotator Tendency Learning. *arXiv preprint arXiv:2503.15237* (2025).

[35] Liyun Zhang, Nanyan Liu, Yuanbin Hou, and Xiaojian Liu. 2014. Uneven illumination image segmentation based on multi-threshold S-F. *Opto-Electronic Engineering* 41, 7 (2014), 81–87.

[36] Liyun Zhang, Zhaojie Luo, Shuqiong Wu, and Yuta Nakashima. 2024. MicroEmo: Time-Sensitive Multimodal Emotion Recognition with Subtle Clue Dynamics in Video Dialogues. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*. 110–115.

[37] Liyun Zhang, Photchara Ratsamee, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. 2023. Panoptic-level image-to-image translation for object recognition and visual odometry enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 2 (2023), 938–954.

[38] Liyun Zhang, Photchara Ratsamee, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. 2022. Thermal-to-Color Image Translation for Enhancing Visual Odometry of Thermal Vision. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 33–40.

[39] Liyun Zhang, Photchara Ratsamee, Bowen Wang, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. 2023. Panoptic-aware image-to-image translation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 259–268.

[40] Le Zhang, Ryutaro Tanno, Moucheng Xu, Yawen Huang, Kevin Bronik, Chen Jin, Joseph Jacob, Yefeng Zheng, Ling Shao, Olga Ciccarelli, et al. 2023. Learning from multiple annotators for medical image segmentation. *Pattern Recognition* 138 (2023), 109400.

[41] Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. 2023. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1977–1987.