SimLabel: Similarity-Weighted Semi-supervision for Multi-annotator Learning with Missing Labels

Liyun Zhang[1], Zheng Lian[2], Hong Liu[3*], Takanori Takebe[4], Yuta Nakashima[5]

[1]Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan
[2]National Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai, China
[3]School of Informatics, Xiamen University, Fujian, China
[4]Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati OH, USA
[5]Institute of Scientific and Industrial Research, The University of Osaka, Osaka, Japan
liyun.zhang@lab.ime.cmc.osaka-u.ac.jp

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. Extensive experiments validate the effectiveness of our method.

Code & Dataset — https://github.com/zly9120/SimLabel
Extended version — https://arxiv.org/abs/2504.09525

Introduction

Multi-annotator learning (MAL) has recently emerged as a research hotspot due to its relevance in subjective or non-deterministic tasks, such as medical diagnosis (Liao et al. 2024), visual perception (Zhang et al. 2023a), etc.. MAL aims to model annotator-specific labeling patterns (Herde, Huseljic, and Sick 2023; Zhang et al. 2025b).

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 (Paolacci, Chandler, and Ipeirotis 2010). 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.

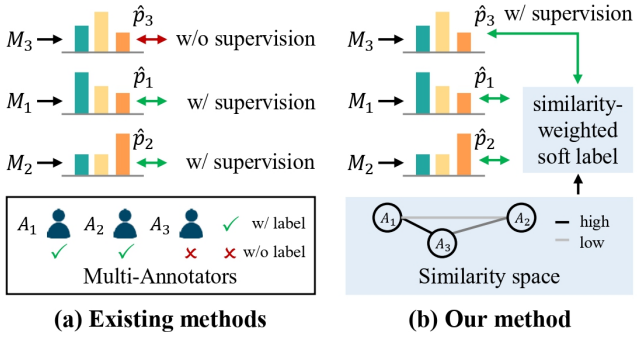

Figure 1: The sample is labeled by annotators A_1 to A_3, with A_3's label missing. (a) In existing methods, the predicted label distribution $\hat{p}_3$ from A_3's model M_3 lacks supervision due to the missing label, resulting in skipped parameter updates for M_3 on this sample. (b) In contrast, our 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.

To this end, we propose a novel similarity-weighted semi-supervised learning framework (SimLabel) to estimate pairwise inter-annotator similarities and leverage them to generate weighted soft labels for missing annotations. Unlike aggregation-oriented similarity with intermediate modeling in Learning From Crowds (LFC) (Jung and Lease 2012; Wu et al. 2024), our similarity iteratively refines to preserve individual patterns. Our goal is not to "fix" inherent missing label characteristics of crowdsourced data, but to enable effective data utilization, ensuring annotator-specific model parameters can be updated for missing labels, not skipped,

*Corresponding author.

improving 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 (Viera and Garrett 2005) 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 $\bar{p}_3$ 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.

Related Work

Traditional Multi-annotator Learning

Traditional multi-annotator learning (MAL) aims to estimate a single consensus or ground-truth label from multiple annotator inputs. Early approaches include probabilistic models (Dawid and Skene 1979), EM algorithms (Raykar et al. 2010; Whitehill et al. 2009), and neural network-based methods (Albarqouni et al. 2016a; Tanno et al. 2019a; Cao et al. 2019). Recent methods employ probabilistic frameworks to aggregate annotations using confusion matrices (Tanno et al. 2019b), agreement distributions (Wang et al. 2023), and Gaussian distributions (Liao et al. 2024). However, this aggregation paradigm treats annotator disagreements as noise to be averaged away rather than valuable information reflecting legitimate differences in annotation patterns (Yan et al. 2014a; Jinadu et al. 2023).

Multi-annotator Labeling Pattern Modeling

Some studies have attempted to model individual annotator patterns rather than aggregating them, including training annotator-specific models on filtered subsets (Mirikharaji et al. 2021), jointly optimizing ground truth and annotator models (Herde, Huseljic, and Sick 2023), associating kernels with prototype libraries (Cheng et al. 2023), and leveraging annotator explanations (Zhang et al. 2023b; Schaekermann et al. 2019). Notably, QuMATL (Zhang et al. 2025c) introduced a paradigm focus shift by modeling individual annotator patterns via learnable queries, viewing each annotator as having unique patterns worth preserving rather than as noisy approximations of ground-truth. However, these methods face a critical challenge: when labels are missing, annotator-specific model parameters cannot be updated, limiting the effectiveness of individual annotator modeling.

Multi-annotator Learning with Missing Labels

To the best of our knowledge, multi-annotator learning with missing labels has not been systematically investigated. While several related works exist (Yan et al. 2014b; Davani, Díaz, and Prabhakaran 2022; Li et al. 2021; Tanno et al. 2019c; Guan et al. 2018; Shah and Zhou 2016; Rodrigues and Pereira 2018; Albarqouni et al. 2016b), they either assume complete annotations, focus on noise handling, or aggregate available labels without addressing missing annotation challenges. Existing MAL methods simply skip annotator-specific model parameter updates when labels are missing, leading to inefficient data utilization and potential overfitting, as the annotator model is trained on limited annotations due to extensive missing labels. Our work fills this critical gap by leveraging inter-annotator similarities to enable continuous parameter updates rather than skipping missing annotations, thereby improving data utilization efficiency and reducing overfitting risks.

Dataset Construction

This paper introduces AMER2, a new multimodal multi-annotator dataset for video emotion recognition (Lian et al.

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

Table 1: Missing rates of the AMER2 dataset (with 10 annotators A_k, $k = 1, \ldots, 10$) are compared to the AMER dataset (with 13 annotators A_k, $k = 1, \ldots, 13$). For each annotator, the missing rate (%) is reported, as well as the average data.

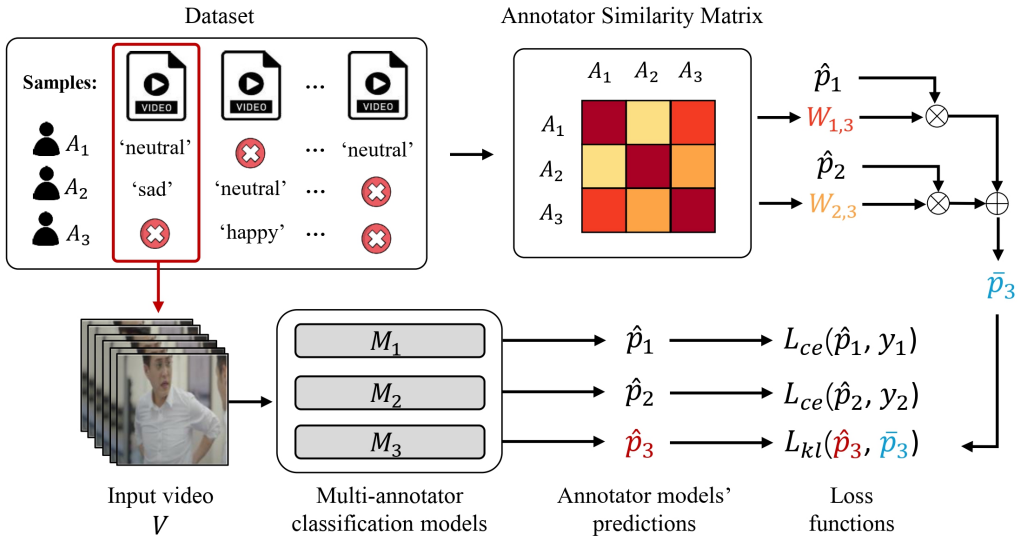

Figure 2: We propose a similarity-weighted semi-supervised learning framework, where annotator similarity is computed using Cohen's kappa (darker indicates higher similarity). For a sample V missing label from A_3, annotator-specific models produce label distributions: labeled predictions ($\hat{p}_1$, $\hat{p}_2$) use cross-entropy loss, while the unlabeled prediction ($\hat{p}_3$) is trained with KL divergence against a soft label $\bar{p}_3$ formed by weighting $\hat{p}_1$ and $\hat{p}_2$ using $W_{1,3}$ and $W_{2,3}$.

2023; Zhang et al. 2024), which extends the existing AMER dataset (Zhang et al. 2025c) with 2,311 additional video samples from movies and TV series. AMER contains 5,207 samples with dense per-annotator labels; its relatively uniform missing rate distribution among the first 10 annotators with sparse annotations (ranging from 79.0% to 81.4%, as shown in Table 1) is insufficient to reflect real-world sparse annotation scenarios and enable evaluation under varying levels of label sparsity. In contrast, Table 1 shows that AMER2 naturally exhibits higher and more variable missing rates among its 10 annotators (ranging from 75.9% to 91.3%), better reflecting real-world sparse annotation challenges and enabling comprehensive evaluation under varying levels of missing labels. Label statistics are provided in the supplementary material (Zhang et al. 2025d).

During the annotation process, we utilize the Label Studio toolkit (Tkachenko et al. 2020-2025) and hire 10 graduate student annotators from our labs. To ensure annotation quality, annotators first undergo preliminary exams on 10 sample videos, selecting the most likely label from 8 emotion categories: *worry*, *happiness*, *neutral*, *anger*, *surprise*, *sadness*, *other*, and *unknown*. These categories are the standard affective computing categories, consistent with AMER for experimental comparability. The samples used in this study underwent rigorous annotation procedures conducted by five affective computing specialists. Each expert independently evaluated the emotional expressions using standardized assessment protocols, achieving consistent annotations of the characters' emotional states.

To maintain annotation quality, annotators failing the preliminary exam are removed from the annotation pool. Each retained annotator completed the annotation task over approximately two weeks with scheduled breaks, providing 201 to 557 labels per annotator. This annotation protocol ensures high-quality labels for reliable evaluation.

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).

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

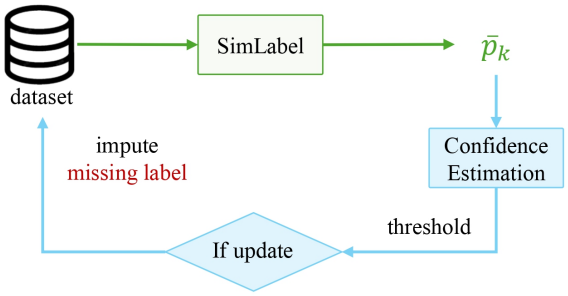

Figure 3: A confidence-based iterative refinement in which confidence for the soft label $\bar{p}_k$ guides pseudo-label imputation. When confidence exceeds a threshold, the pseudo-label fills the missing label and updates the inter-annotator similarity matrix for subsequent soft-label generation.

$\mathcal{Y} = \{y\}$ 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 (Liao et al. 2024)), confusion matrix (MaDL (Herde, Huseljic, and Sick 2023)), or query-based architecture (QuMATL (Zhang et al. 2025c)), 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 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 (Viera and Garrett 2005) 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.

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

Require: Dataset $\mathcal{D}$ with missing labels, confidence threshold T

Initialize similarity matrix SM using Cohen's kappa coefficient on available labels
Initialize annotator models $\{M_1, M_2, \ldots, M_K\}$
for each training epoch **do**
  for each sample with missing labels **do**
    Generate similarity-weighted soft label: $\bar{p}_k = \sum w_{k',k} \hat{p}_{k'}$
    Calculate confidence: $c = \max(\bar{p}_k) \times (1 - \frac{H[\bar{p}_k]}{H_{\max}})$
    if $c \geq T$ **then**
      Extract predicted label: $y_{pred} = \arg\max(\bar{p}_k)$
      Impute y_{pred} into dataset for corresponding missing annotation
      Recalculate annotator similarity matrix SM using updated dataset
    end if
    Update annotator models using supervised and semi-supervised losses
  end for
end for

Ensure: Refined similarity matrix SM, imputed dataset, trained annotator models

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 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:

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

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 $H_{\text{max}} = \log N$. 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 $H_{norm}(\bar{p}_k) = (1 - \frac{H[\bar{p}_k]}{H_{\text{max}}})$ 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.

Experiment

We conduct experiments comparing SimLabel (with and without confidence-based iterative refinement) against existing approaches that skip annotator-specific model parameter updates for missing labels. We evaluate representative modeling frameworks: Gaussian distribution fitting (PADL (Liao et al. 2024)), confusion matrix (MaDL (Herde, Huseljic, and Sick 2023)), and query-based modeling (QuMATL (Zhang et al. 2025c)). Experiments are conducted on AMER2 and AMER (real missing labels), and STREET dataset (simulated missing labels), using Accuracy and Difference of Inter-annotator Consistency (DIC) (Zhang et al. 2025c) as evaluation metrics. Note that we discuss additional key issues with experimental results, including training dynamics with failure-case analysis on utilization-accuracy tradeoff, threshold sensitivity to missing rates and cross-domain datasets, and strategies for handling noisy labels and avoiding propagation errors, etc., in the supplementary material (Zhang et al. 2025d).

Implementation Details

We use Cohen's kappa coefficient (Viera and Garrett 2005) 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.

Evaluation Metrics and Datasets

For evaluation metrics, Accuracy is a standard metric to evaluate individual annotator modeling. DIC (Zhang et al. 2025a) 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.

For datasets, we utilize a contributed AMER2, an earlier version AMER, and a city impression dataset STREET (Zhang et al. 2025c). AMER2 and AMER contain real-world missing labels. STREET is a complete dataset, we randomly remove labels at different ratios to simulate missing labels. We also apply random removal to AMER2 and AMER to further increase missing rates.

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.

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 2: Accuracy comparison on AMER2 dataset (10 annotators, A_k, k = 1, ..., 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}	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 3: Accuracy comparison on AMER dataset evaluating annotator modeling performance for 13 annotators.

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**

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.

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.

For the gain evaluation of inter-annotator consistency, the DIC scores in Table 5 also show that our similarity-weighted

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**

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.

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**

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.

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.

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 (Viera and Garrett 2005) and self-iterative framework, the model performance demonstrates robustness. Second, we conducted ablation experiments comparing Cohen's kappa with Pearson correlation coefficient (Benesty et al. 2009) and Krippendorff's alpha coefficient (Gwet 2011) 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.

Conclusion

We addressed missing labels in multi-annotator learning through a similarity-weighted semi-supervised framework that leverages inter-annotator relationships rather than skipping updates for annotators without labels. SimLabel integrates soft-label generation with a confidence-based iterative refinement mechanism to adaptively update similarity estimates. We also introduce AMER2, a new dataset with high and variable missing rates that reflects real-world annotation sparsity. Experiments across multiple datasets and sparsity levels demonstrate the effectiveness of SimLabel. In future work, we plan to extend the framework to dynamic annotator behaviors and more complex annotation scenarios.

Acknowledgments

This work was supported by JST Grant Number JPMJPF2115 and the Future Social Value Co-Creation Project, the University of Osaka. This work is also supported by the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (No. MAIS2024311) and Youth Science Fund Project of National Natural Science Foundation of China (No. 62201572).

References

Albarqouni, S.; Baur, C.; Achilles, F.; Belagiannis, V.; Demirci, S.; and Navab, N. 2016a. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5): 1313–1321.

Albarqouni, S.; Baur, C.; Achilles, F.; Belagiannis, V.; Demirci, S.; and Navab, N. 2016b. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5): 1313–1321.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, 1–4. Springer.

Cao, P.; Xu, Y.; Kong, Y.; and Wang, Y. 2019. Max-mig: an information theoretic approach for joint learning from crowds. *arXiv preprint arXiv:1905.13436*.

Cheng, Y.-C.; Shiau, Z.-Y.; Yang, F.-E.; and Wang, Y.-C. F. 2023. TAX: Tendency-and-Assignment Explainer for Semantic Segmentation with Multi-Annotators. *arXiv preprint arXiv:2302.09561*.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

Guan, M.; Gulshan, V.; Dai, A.; and Hinton, G. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Gwet, K. L. 2011. On the Krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October*, 2(2011): 2011.

Herde, M.; Huseljic, D.; and Sick, B. 2023. Multi-annotator Deep Learning: A Probabilistic Framework for Classification. *arXiv preprint arXiv:2304.02539*.

Jinadu, U.; Annan, J.; Wen, S.; and Ding, Y. 2023. Loss Modeling for Multi-Annotator Datasets. *arXiv preprint arXiv:2311.00619*.

Jung, H. J.; and Lease, M. 2012. Improving Quality of Crowdsourced Labels via Probabilistic Matrix Factorization. In *HCOMP@ AAAI*.

Li, J.; Sun, H.; Li, J.; Chen, Z.; Tao, R.; and Ge, Y. 2021. Learning from multiple annotators by incorporating instance features. *arXiv preprint arXiv:2106.15146*.

Lian, Z.; Sun, H.; Sun, L.; Chen, K.; Xu, M.; Wang, K.; Xu, K.; He, Y.; Li, Y.; Zhao, J.; et al. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9610–9614.

Liao, Z.; Hu, S.; Xie, Y.; and Xia, Y. 2024. Modeling annotator preference and stochastic annotation error for medical image segmentation. *Medical Image Analysis*, 92: 103028.

Mirikharaji, Z.; Abhishek, K.; Izadi, S.; and Hamarneh, G. 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.

Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5): 411–419.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning From Crowds. *Journal of Machine Learning Research*, 11(43): 1297–1322.

Rodrigues, F.; and Pereira, F. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Schaekermann, M.; Beaton, G.; Habib, M.; Lim, A.; Larson, K.; and Law, E. 2019. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

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

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019a. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019b. 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.

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019c. 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.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Viera, A.; and Garrett, J. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine,Family Medicine*.

Wang, C.; Gao, Y.; Fan, C.; Hu, J.; Lam, T. L.; Lane, N. D.; and Bianchi-Berthouze, N. 2023. Learn2agree: Fitting with multiple annotators without objective ground truth. In *International Workshop on Trustworthy Machine Learning for Healthcare*, 147–162. Springer.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J.; and Ruvolo, P. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22.

Wu, X.; Jiang, L.; Zhang, W.; and Li, C. 2024. Worker Similarity-based Label Completion for Crowdsourcing. *IEEE Transactions on Big Data*.

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014a. Learning from multiple annotators with varying expertise. *Machine learning*, 95: 291–327.

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014b. Learning from multiple annotators with varying expertise. *Machine learning*, 95: 291–327.

Zhang, L.; Ke, J.; Fan, S.; Sha, X.; and Lian, Z. 2025a. A Unified Evaluation Framework for Multi-Annotator Tendency Learning. *arXiv preprint arXiv:2508.10393*.

Zhang, L.; Lian, Z.; Liu, H.; Takebe, T.; and Nakashima, Y. 2025b. QuMAB: Query-based Multi-Annotator Behavior Modeling with Reliability under Sparse Labels. *arXiv preprint arXiv:2507.17653*.

Zhang, L.; Lian, Z.; Liu, H.; Takebe, T.; and Nakashima, Y. 2025c. QuMATL: Query-based Multi-annotator Tendency Learning. *arXiv preprint arXiv:2503.15237*.

Zhang, L.; Lian, Z.; Liu, H.; Takebe, T.; and Nakashima, Y. 2025d. SimLabel: Similarity-Weighted Semi-supervision for Multi-annotator Learning with Missing Labels. *arXiv preprint arXiv:2504.09525*.

Zhang, L.; Luo, Z.; Wu, S.; and Nakashima, Y. 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.

Zhang, L.; Tanno, R.; Xu, M.; Huang, Y.; Bronik, K.; Jin, C.; Jacob, J.; Zheng, Y.; Shao, L.; Ciccarelli, O.; et al. 2023a. Learning from multiple annotators for medical image segmentation. *Pattern Recognition*, 138: 109400.

Zhang, Y.; Gu, S.; Gao, Y.; Pan, B.; Yang, X.; and Zhao, L. 2023b. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1977–1987.