IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

L2CU: Learning To Complement Unseen Users

DILEEPA PITAWELA[1], GUSTAVO CARNEIRO[2], HSIANG-TING CHEN[1]
[1]The School of Computer Science, University of Adelaide, Australia
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Corresponding author: Hsiang-Ting Chen (tim.chen@adelaide.edu.au)

ABSTRACT Recent research highlights the potential of machine learning models to learn to complement (L2C) human strengths; however, generalizing this capability to unseen users remains a significant challenge. Existing L2C methods oversimplify interaction between human and AI by relying on a single, global user model that neglects individual user variability, leading to suboptimal cooperative performance. Addressing this, we introduce L2CU, a novel L2C framework for human-AI cooperative classification with unseen users. Given sparse and noisy user annotations, L2CU identifies representative annotator profiles capturing distinct labeling patterns. By matching unseen users to these profiles, L2CU leverages profile-specific models to complement the user and achieve superior joint accuracy. We evaluate L2CU on datasets (CIFAR-10N, CIFAR-10H, Fashion-MNIST-H, Chaoyang and AgNews), demonstrating its effectiveness as a model-agnostic solution for improving human-AI cooperative classification.

INDEX TERMS Human-AI Cooperation, Learning To Complement

I. INTRODUCTION

HUMAN-AI cooperation aims to combine the strengths of humans and AI to achieve superior performance compared to either acting alone. Within this field, learning to defer (L2D) and learning to complement (L2C) represent distinct approaches. L2D focuses on AI abstention when confidence is low, relying on human intervention for difficult cases [1]–[3]. L2C, however, aims for a more synergistic partnership, where both human and AI actively contribute their complimentary strengths, leading to greater overall performance [4], [5]. Both L2D and L2C are vital strategies for effective human-AI cooperative tasks. L2D prioritizes efficient use of human time by focusing human expertise on instances where AI is uncertain, whereas L2C priorities maximal joint accuracy through the human-AI cooperation.

A key challenge in deploying human-AI cooperative systems is ensuring its effective generalization to diverse skills and behaviors of unseen users (i.e., drawn from the same distribution as users in the training set but not included in it). Recent advances in L2D begin to address this challenge. For example, EA-L2D [6] employs a Bayesian approach for expert-agnostic modeling and generalizes to unseen experts. L2D-Pop [7] utilizes meta-learning to adapt to new experts. In contrast, the challenge of complementing unseen users in L2C remains largely under-explored (Figure 1). The task is inherently more complex than L2D's binary deferral because L2C requires adapting a model to latent user characteristic labeling patterns based on sparse and potentially incorrect labels, given that such sparsity and noise are inevitable in real-world datasets.

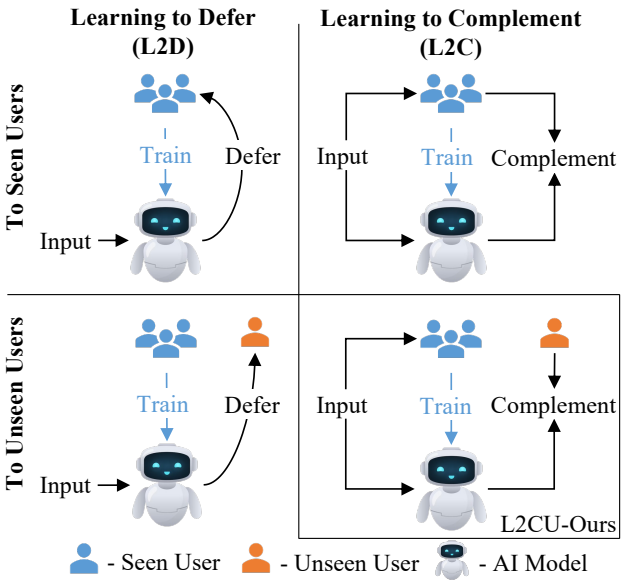

FIGURE 1. Paradigms of Human-AI Cooperation with seen users in blue and unseen users in orange. L2D defers decisions to humans, evolving to handle both seen and unseen users. L2C complements human strengths, and L2CU (Ours) advances L2C to complement unseen users.

To address this research gap, this paper introduces L2CU,

a novel L2C framework for human-AI cooperative classification, designed to achieve optimal performance with unseen users (Figure 2). Given a sparse training dataset with noisy labels, where training users only label subset of samples, L2CU extracts representative annotator profiles, each capturing a distinct noisy labeling pattern. These profiles are then used to train AI cooperative model instances, each optimized for a specific profile. To enable cooperation at test time, a user profiling process matches a new user to a profile, and the corresponding AI cooperative model is selected. This profiling mechanism enables L2CU to generalize to unseen users who were not included in training, and adapt to their characteristic labeling patterns.

We thoroughly evaluate L2CU on both simulated and real multi-rater settings across diverse modalities (image and text) and domains (everyday objects, news classification, and medical diagnosis), with CIFAR-10N, CIFAR-10H, Fashion-MNIST-H, Chaoyang and AgNews datasets. We also introduce a novel assessment metric, alteration rate, which quantifies the extent to which the model's predictions improve upon or deviate from the original human labels. Our results demonstrate that L2CU, a model-agnostic L2C framework, generalizes effectively to unseen users, consistently outperforming both individual human annotators and leading L2D and L2C methods across a range of classification tasks. Our key contributions include:

- L2CU, a new learning-to-complement framework designed to complement users unseen during training.
- L2CU handles sparse, multi-user settings and proposes a label augmentation method to augment sparse training data while preserving characteristic labeling patterns.
- L2CU achieves leading performance in human-AI cooperative classification and introduces a novel alteration rate metric to offer insights into the model's impact on human labels.

L2CU's model-agnostic design, ability to train using sparse noisy labels (without accessing ground truth), and generalization to unseen users make it a significant contribution to the field of human-AI cooperative classification

II. RELATED WORK

The uncertainties of automation often demand human involvement, leading to new human-AI cooperation paradigms [8]. Learning-to-Defer (L2D) methods let AI models handle confident cases while deferring uncertain ones to humans by optimizing a utility function that balances model accuracy, preference of a human decision, and deferral costs [1]–[3]. For instance, [9] used AI model ensemble to flag high-risk patients for human review. [10] refines classifiers and use post-hoc rejectors to identify when to defer. [2], [11], [12] optimize surrogate loss functions for deferral. [13] optimizes work distribution for deferral. However, these assume clean labels or defer only to seen users. In contrast, we propose a learning-to-complement approach that removes the need for clean labels and supports cooperation with unseen users.

Learning-to-Defer to Unseen Users (L2DU) aims to defer to test users different from the training users. EA-L2D [6] proposes a Bayesian framework to model expert behavior without expert-specific training data. With this prior knowledge about experts, they generalize to unseen experts. L2D-Pop [7] uses a small context set to characterize an expert's decision patterns and employs meta-learning to refine deferral policies dynamically. However, it assumes each training user labels all training samples—uncommon in multi-rater settings. In contrast, being a L2C method, L2CU addresses realistic conditions where each training user labels only a subset of the data.

Learning-to-Complement (L2C) is less explored than L2D. L2C leverages the strengths of both humans and AI to improve decision-making. [4] considers uncertainty from both sides to improve decision, [5] uses Bayesian modeling for human-AI complementary. LECOMH [14] estimates human-AI consensus, then trains a selection module to minimize error and collaboration cost. LECODU [15] further decides when to collaborate or defer, and how many experts to involve. However, unlike those methods that complement the same users appearing in training, L2CU enables complementing unseen users.

While adapting to users is studied in recommender systems, the absence of user-item preference data and L2C's distinct goal of complementing user biases to improve joint accuracy—unlike predicting preferences—render such methods inapplicable, leaving user adaptation in L2C underexplored [16], [17]. Available L2C work model users globally using confusion matrices [18] or behavior models [19], but these overlook individual biases and require extensive human labels. In contrast, L2CU identifies distinct annotator profiles and adapts without requiring extra labels to train behavior models. Appendix F extends the related work.

III. METHODOLOGY

A. PROBLEM FORMULATION

We consider a classification problem with C classes, where each data sample has multiple, potentially noisy labels provided by different annotators. Our training dataset, denoted as $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \{\tilde{\mathbf{y}}_{i,j}\}_{j\in\mathcal{A}})\}_{i=1}^{N}$, consists of N data samples. Each sample $\mathbf{x}_i \in \mathcal{X}$ has a set of noisy labels $\{\tilde{\mathbf{y}}_{i,j}\}_{j\in\mathcal{A}}$, where $\tilde{\mathbf{y}}_{i,j} \in \{0, 1, ..., C-1\}$ represents the label provided by annotator j for sample i. The set $\mathcal{A}$ represents the set of all annotators. Crucially, each annotator labels only a *subset* of the samples. We assume each sample $\mathbf{x}_i$ has a latent, clean label $\mathbf{y}_i$, and that annotator noise is class-dependent [20]. When clean labels are unavailable, we employ Crowdlab [21] to estimate consensus labels, $\bar{\mathbf{y}}_i$, for each sample, resulting in a consensus-labeled dataset $\bar{\mathcal{D}} = \{(\mathbf{x}_i, \bar{\mathbf{y}}_i)\}_{i=1}^{N}$. Appendix A provides more details on consensus estimation.

B. TRAINING OF HUMAN-AI COOPERATIVE MODEL

To enable complementing unseen users given sparse, noisy multi-rater data, L2CU's training (Fig.2) proceeds as follows: First, a set of representative annotator profiles are identified,

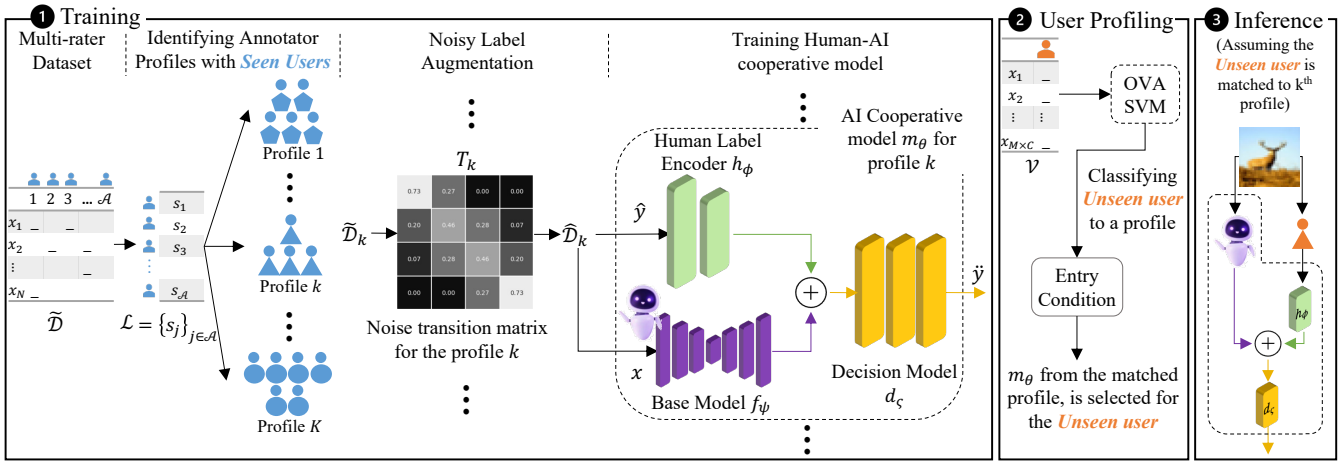

FIGURE 2. Three step L2CU framework. 1) During training, from a sparse multi-rater dataset, unique annotator profiles are identified (1, ..., *K*). Then, for each annotator profile, noisy label augmentation is performed and a AI cooperative model is trained. 2) During user profiling, a test user annotates a validation set and based on validation labels, a profile is matched by OVA SVM, entry condition is evaluated, and respective AI cooperative model is selected. 3) At inference, the test user is paired with the corresponding model from the selected profile for cooperative classification.

each capturing distinct noisy labeling patterns among users. Next, noisy labels are augmented for each profile, to mitigate data sparsity. Finally, AI cooperative model instances are trained for each profile using the augmented data, enabling cooperation and effective generalization to unseen users who can be matched to a profile. We explain each step in more details below.

1) Identifying Annotator Profiles

In multi-rater datasets, annotators often exhibit similar labeling error patterns, reflecting individual biases or areas of confusion [20], [22]. We define these characteristic error patterns as *annotator profiles*. For example, on CIFAR-10, an annotator profile might be characterized by frequent misclassifications of horses as deer, birds as planes, and trucks as automobiles [22].

To identify representative annotator profiles, we first construct a fixed-length *label vector* $\mathbf{s}_j$ for each annotator $j \in \mathcal{A}$. This vector represents the annotator's characteristic labeling pattern across all C classes. The construction of $\mathbf{s}_j$ proceeds as follows: 1) For each class $c \in \{1, ..., C\}$, we gather all noisy labels $\tilde{\mathbf{y}}_{i,j}$ provided by annotator j for samples with consensus label c. This forms a set of labels for each class: $\mathcal{S}_j^{(c)} = \{\tilde{\mathbf{y}}_{i,j} | (\mathbf{x}_i, \tilde{\mathbf{y}}_{i,j}) \in \tilde{\mathcal{D}}\}$. 2) From each set $\mathcal{S}_j^{(c)}$, we randomly select L labels. If an annotator has provided fewer than L labels for a given class, that annotator is excluded from further analysis. The value of L is chosen to balance profile representation with the inclusion of as many annotators as possible; in our experiments, L is set to 20. 3) The selected noisy labels are then concatenated across all C classes to form the label vector $\mathbf{s}_j$ where each $l_i^{(c)}$ is a noisy label $\tilde{\mathbf{y}}_{i,j}$ selected from $\mathcal{S}_j^{(c)}$.

$$\mathbf{s}_j = [\underbrace{l_1^{(1)}, ..., l_L^{(1)}}_{\text{Class 1}}, \underbrace{l_1^{(2)}, ..., l_L^{(2)}}_{\text{Class 2}}, ..., \underbrace{l_1^{(C)}, ..., l_L^{(C)}}_{\text{Class C}}] \quad (1)$$

This process yields a set of label vectors, $\mathcal{L} = \{\mathbf{s}_j\}_{j \in \mathcal{A}}$. Although the specific labels within each $\mathbf{s}_j$ will vary, the consistent class ordering allows for direct comparison of labeling patterns across annotators.

To identify distinct annotator profiles, we cluster the label vectors in $\mathcal{L}$ using Fuzzy K-Means, chosen for its robustness to noise in the label vectors [23], [24]. The number of clusters, K, is determined using the silhouette score (see Appendix B), which provides a measure of cluster cohesion and separation. Each annotator is then assigned to the profile (cluster) to which their label vector has the highest membership score.

2) Noisy-label Augmentation

Because each annotator labels only a subset of the data, and annotators are grouped into profiles, the original training set $\tilde{\mathcal{D}}$ is now divided into profile-specific subsets $\{\tilde{\mathcal{D}}_k\}_{k=1}^K$ (where $\tilde{\mathcal{D}}_k$ contains labels from users in profile k) resulting in sparse data for training each profile-specific model. To address this data sparsity while preserving profile-specific noise patterns, we perform noisy label augmentation on each profile, enabling the model to learn these patterns during training.

This augmentation process leverages the characteristic noisy labeling patterns captured by each profile. Specifically, for each profile k, we estimate a *label transition matrix* $\mathbf{T}_k \in [0, 1]^{C \times C}$. This matrix quantifies the probability of annotators within profile k assigning a noisy label n to a sample having consensus label c, effectively capturing the profile specific biases. The transition matrix is estimated as,

$$\mathbf{T}_k(c, n) = \frac{1}{|\mathcal{S}_k^{(c)}|} \sum_{\tilde{\mathbf{y}}_i \in \mathcal{S}_k^{(c)}} \mathbb{I}[\tilde{\mathbf{y}}_i = n] \quad (2)$$

where $\mathcal{S}_k^{(c)} = \bigcup_{j \in \mathcal{A}_k} \mathcal{S}_j^{(c)}$ represents the set of noisy labels provided by annotators in profile k ($\mathcal{A}_k \subset \mathcal{A}$) for samples with consensus label c, and $\mathbb{I}[\cdot]$ is the indicator function.

New noisy labels are then generated for each profile by sampling from this profile-specific transition matrix $\mathbf{T}_k$. This augmentation process ensures sufficient training data for each profile-specific model while preserving the characteristic noise patterns of that profile, allowing the model to learn to effectively complement the users within that profile.

Each element $\mathbf{T}_k(c,n)$ of the transition matrix (Eq. 2) represents the probability, $P(\tilde{Y}=n|\bar{Y}=c,R=k)$, that an annotator in profile k ($R=k$) assigns the noisy label $\tilde{Y}=n$ to a sample with consensus label $\bar{Y}=c$. To augment the data for profile k, we proceed as follows: For each data sample $\mathbf{x}_i$ in the profile's subset $\tilde{\mathcal{D}}_k$, we retrieve its consensus label c from $\bar{\mathcal{D}}$. Using the c-th row of the transition matrix $\mathbf{T}_k$, which represents the categorical distribution over noisy labels for class c, we sample G new noisy labels, $\{\hat{\mathbf{y}}_{i,g}\}_{g=1}^{G}$. This results in an augmented training set for profile k: $\hat{\mathcal{D}}_k = \{(\mathbf{x}_i, \{\hat{\mathbf{y}}_{i,g}\}_{g=1}^{G})\}_{i=1}^{N}$.

3) Training Human-AI Cooperative Model

To effectively complement unseen users by leveraging learned annotator profiles and their associated augmented noisy labels, we introduce the AI cooperative model architecture (final step of Training, Fig. 2). The model consists of three components. 1) A base model, $f_{\psi_k}: \mathcal{X} \rightarrow \mathbb{R}^C$, extracts features from the input data, transforming it into a logit vector. 2) A human label encoder, $h_{\phi_k}: \mathcal{Y} \rightarrow \mathbb{R}^C$, models the profile-specific noisy labeling patterns. Finally, 3) a decision model, $d_{\zeta_k}: \mathbb{R}^C \times \mathbb{R}^C \rightarrow \Delta^{C-1}$ learns the joint noise distribution of f_{ψ_k} and h_{ϕ_k} and produces a categorical distribution across classes where we use the one with highest probability as the final prediction. The whole model $m_{\theta_k}: \mathcal{X} \times \mathcal{Y} \rightarrow \Delta^{C-1}$ is defined as:

$$m_{\theta_k}(\mathbf{x}, \hat{\mathbf{y}}) = d_{\zeta_k}(f_{\psi_k}(\mathbf{x}) \oplus h_{\phi_k}(\hat{\mathbf{y}})), \tag{3}$$

where $\theta_k = \{\psi_k, \phi_k, \zeta_k\}$, and $\oplus$ represents the concatenation operator. Note that, for each profile k, we train a single m_{θ_k} using profile's $\hat{\mathcal{D}}_k$. The base model $f_{\psi_k}(.)$ could use any architecture. Similarly, $h_{\phi_k}(.)$ and $d_{\zeta_k}(.)$ can be of different architectures; we configured them as a two-layer and three-layer multi-layer perceptron, respectively, with ReLU activations. The model in eq.3 is trained with:

$$\begin{aligned}\{\theta_k^*\}_{k=1}^{K} = \arg\min_{\{\theta_k\}_{k=1}^{K}} \frac{1}{K \times |\hat{\mathcal{D}}_k| \times G} \times \\ \sum_{k=1}^{K} \sum_{(\mathbf{x}_i, \{\hat{\mathbf{y}}_{i,g}\}_{g=1}^{G}) \in \hat{\mathcal{D}}_k} \ell\left(\bar{\mathbf{y}}_i, m_{\theta_k}(\mathbf{x}_i, \hat{\mathbf{y}}_{i,g})\right) + \\ \lambda \times \ell\left(\hat{\mathbf{y}}_{i,g}, (\mathbf{T}_k)^{\top} \times m_{\theta_k}(\mathbf{x}_i, \hat{\mathbf{y}}_{i,g})\right),\end{aligned} \tag{4}$$

where $\bar{\mathbf{y}}_i$ is the consensus label from $\bar{\mathcal{D}}$, $\ell(.)$ is the cross-entropy loss, $\lambda \in [0, \infty]$ is a hyper-parameter, and the second loss term is motivated by the forward correction procedure proposed by [25], transforming the clean label prediction from $m_{\theta_k}(.)$ into the noisy ones in $\hat{\mathcal{D}}_k$.

C. USER PROFILING AND INFERENCE

At test time, an unseen user undergoes a *user profiling* process (User Profiling phase of Fig. 2) to determine the most matching AI cooperative model instance, m_{θ_k}, for classification. This process consists of two key steps: 1) Assigning the user to one of the K annotator profiles identified during training. 2) Evaluating an entry condition to determine whether the cooperative model instance $m_{\theta_k}(\cdot)$ from the assigned profile should be used.

In the first step, the test user annotates a small validation set $\mathcal{V} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M \times C}$. This validation set contains M samples from each C class, does not overlap with the training or testing sets, and has clean labels. A key advantage of L2CU is that it only needs $M \times C$ samples to adapt to unseen users—unlike prior methods that require extensive expert annotations to train additional behavioral models. From the obtained labels from the test user, we construct a label vector having the same format as Eq. 1 and classify the user into a profile identified during training phase using a one-versus-all (OVA) support vector machine (SVM) classifier. OVA SVM takes the test user's label vector and outputs a categorical distribution across profiles where we choose the one with highest prediction as the matching profile. This empirically leads to better performance compared to soft assignment (see Ablation Sec.VI).

The OVA SVM classifier is trained on label vectors collected from users belong to identified profiles during the training. Specifically, user labels for M samples from each of the C classes (using consensus labels to determine the class), are randomly collected, formatted as in (Eq. 1), labeled with their corresponding profile, and used for training. Given their strong performance in high dimensions with limited data, SVMs are well-suited for this task.

In the second step, building on [5], the entry condition compares the accuracy of the base model $f_{\psi_k}(.)$ and the testing user on the validation set $\mathcal{V}$. The $m_{\theta_k}(.)$ from the predicted profile is paired with the test user if the base model outperforms the user; else, the user is rejected. Once the test user is paired with a model, they perform cooperative classification (Inference of Fig.2). $m_{\theta_k}(.)$ is evaluated on a test set $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$ having clean labels that does not overlap with training or validation sets.

D. MEASURES FOR H-AI COOPERATIVE CLASSIFICATION

As part of cooperative classification, L2CU alters the user labels to improve accuracy. These alterations can be positive (correcting an incorrect user label) or negative (changing a correct user label to an incorrect one). To capture this, we introduce positive and negative alteration measures,

$$\begin{aligned}&\underset{A_+}{\text{Positive}}\\ &\text{Alteration} = \frac{1}{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\frac{|\mathcal{M}^{\mathbf{c}}{}_j|}{|\mathcal{I}_j|} \quad \underset{A_-}{\text{Negative}}\\ &\qquad\qquad \text{Alteration} = \frac{1}{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\frac{|\mathcal{M}^{\mathbf{e}}{}_j|}{|\mathcal{R}_j|}\end{aligned} \tag{5}$$

where $\mathcal{I}_j = \{i | i \in \mathcal{T} \text{ and } \tilde{\mathbf{y}}_{i,j} \neq \mathbf{y}_i\}$ represents the set of samples incorrectly labeled by the j^{th} user, $\mathcal{M}^{\mathbf{c}}{}_j = \{i | i \in \mathcal{I}_j \text{ and } \ddot{\mathbf{y}}_{i,j} = \mathbf{y}_i\}$ represents the set of samples labeled incorrectly by user j but corrected by the model, $\mathcal{R}_j = \{i | i \in$

$\mathcal{T}$ and $\tilde{\mathbf{y}}_{i,j} = \mathbf{y}_i\}$ represents the set of samples correctly labeled by the j^{th} user, $\mathcal{M}^{e}{}_j = \{i | i \in \mathcal{R}_j \text{ and } \ddot{\mathbf{y}}_{i,j} \neq \mathbf{y}_i\}$ represents the set of samples labeled correctly by user j but later mislabeled by the model. Note that $\mathbf{y}_i$ represents the test set clean label and $\ddot{\mathbf{y}}_{i,j} = \text{Scalar}(m_{\theta_k}(\mathbf{x}_i, \tilde{\mathbf{y}}_{i,j}))$, with the function Scalar returning a scalar label representing the class with the largest prediction from the model $m_{\theta_k}(.)$. In eq.5, A_+ measures the proportion of a user's incorrect labels that the model successfully corrected. In contrast, A_-, in eq.5, measures the proportion of a user's correct labels that were incorrectly altered by the model. In edge cases where the annotator is perfect or always wrong, A_+ and A_- becomes zero respectively, for division by zero.

We also assess original accuracy (before labels alterations) and post-alteration accuracy (after AI cooperative model alters labels). These are computed per user and averaged across all users to determine overall improvement. An effective model should have high A_+, high post-alteration accuracy and low A_-.

TABLE 1. Accuracy of L2CU vs. related methods. Unlike others, L2CU outperforms without training ground truths and by complementing unseen users. (Missing values due to lack of sparse multi-rater support.)

Method	CIFAR -10	CIFAR -10N	CIFAR -10H	F-MNI-ST-H	Chao-yang
Learning To Defer					
MOE [1]	-	0.831	0.812	0.600	0.583
CC [9]	-	0.970	0.971	0.801	0.863
CE [11]	-	0.949	0.967	0.729	0.706
DifT [10]	-	0.940	0.944	0.704	0.765
OVA [12]	-	0.959	0.974	0.794	0.845
WSP [2]	-	0.948	0.976	0.775	0.872
Learning to Defer to Unseen Users					
L2D-Pop [7]	0.947	-	-	-	0.970
EA-L2D [6]	0.820	-	-	-	-
Learning To Complement					
LECOMH [14]	-	-	0.988	-	0.988
LECODU [15]	0.951	-	0.989	-	0.990
L2CU (Ours)	**0.968 ±0.002**	**0.989 ±0.001**	**0.993 ±0.002**	**0.878 ±0.008**	**0.991 ±0.004**

IV. EXPERIMENT SETUP

A. DATASETS

CIFAR-10 [26] includes 50000 training, 200 validation, and 9800 testing images, across 10 classes. **CIFAR-10N** [27] extends CIFAR-10's training set with labels collected from 747 annotators, with each image having three independent labels. **CIFAR-10H** [28] expands CIFAR-10's testing set with labels from 2571 annotators, resulting in an average of 51 labels per image. **Fashion-MNIST-H** [29] extends Fashion-MNIST's [30] testing set with labels from 885 annotators, averaging 66 labels per image. We use this testing set as the training set, with 200 images from the original training set allocated for validation and the remainder for testing. Lastly, **Chaoyang** [31] is a four-class pathological dataset with 4021 training, 80 validation, and 2059 testing images, each having three expert labels in the training set. Unless stated otherwise, above validation sets are composed of randomly selected subsets of samples from the respective original test sets, with selected samples excluded from testing.

B. SETUP ON DATASETS WITH SIMULATED ANNOTATORS

On CIFAR-10, a pairwise flipping experiment is conducted where 8 out of 10 classes have clean labels, but in two classes, 60% of samples have labels flipped. Three user profiles are simulated, one that flips labels between classes airplane↔bird, another profile that flips horse↔deer, and the other flips truck↔automobile. For each profile, five training and five testing users are simulated, resulting in 15 unique users in each set. Labels from training users combined with training samples, form $\tilde{\mathcal{D}}$ from which we identify K profiles, train m_{θ_k} in eq. 3 for each k with a ResNet-18 [32] as $f_{\psi_k}(.)$, and train OVA SVM.

C. SETUP ON DATASETS WITH REAL ANNOTATORS

With CIFAR-10N, we conduct two experiments. In the first experiment, the labels from 747 annotators form $\tilde{\mathcal{D}}$. Of these, 159 annotators who labelled at least 20 images per class ($L = 20$) are selected, split into 79 training users and 80 testing users. The training users' labels are used to identify K profiles where K is chosen from silhouette score, train m_{θ_k} for each profile, and train the OVA SVM. During testing, noisy-label transition matrices are estimated using annotator labels and consensus labels for each test user, resulting in 80 noisy test sets. In the second CIFAR-10N experiment, users in CIFAR-10H are used as a testing set. Noise transition matrices are estimated and used to simulate noisy annotations for each testing user, resulting in unique noisy test sets for all 2571 users. For Fashion-MNIST-H, labels from 885 annotators form $\tilde{\mathcal{D}}$. 366 annotators who labeled at least 20 images per class are selected, split into 183 training and 183 testing sets. Training users are used to identify K profiles, train m_{θ_k}, and train an OVA SVM. During testing noisy-label transition matrices are estimated for each testing user and produced 183 noisy testing sets. Chaoyang dataset has three annotators per image, forming $\tilde{\mathcal{D}}$ which is used to build K profiles, train m_{θ_k}, and train an OVA SVM. During testing, noisy-label transition matrices are estimated, resulting in three noisy test users. More details on experiment setup, data preparation, and implementation are in Appendix C.

D. TRAINING DETAILS

Data augmentation policies in [33], [34] were adopted for CIFAR-10 and Fashion-MNIST respectively while Chaoyang is limited to random resized crops of 224×224. Training runs for 500 epochs with early stopping after 20 unimproved epochs. As found empirically (see Ablation Sec.VI), we keep $\lambda = 0.1$ that yields optimal performance. We use Imagenet1K [35] pre-trained backbone models as base model. Adam and NAdam optimize $f_{\psi_k}(.)$ and $m_{\theta_k}(.)$ respectively, in PyTorch on a NVIDIA RTX 4090.

TABLE 2. Number of users who improved (I), maintained (M), or did not improve (NI) out of the test users profiled and satisfied entry condition. Includes original vs. post-alteration accuracy with positive (A_+) and negative (A_-) alterations along with chosen K from silhouette score.

Dataset	K (Silhouette score)	Test Users	I	M	NI	Original Accuracy	Post. Alt. Accuracy	A_+	A_-
With simulated annotators									
CIFAR-10	3 (0.34)	15	15	0	0	0.880	0.968	0.953	0.09
With real annotators									
CIFAR-10N	2 (0.01)	80	80	0	0	0.836	0.989	0.954	0.004
CIFAR-10H	2 (0.01)	2022	2022	0	0	0.940	0.993	0.939	0.004
F-MNIST-H	2 (0.09)	182	182	0	0	0.662	0.878	0.758	0.073
Chaoyang	3 (0.99)	2	2	0	0	0.858	0.988	0.968	0.045

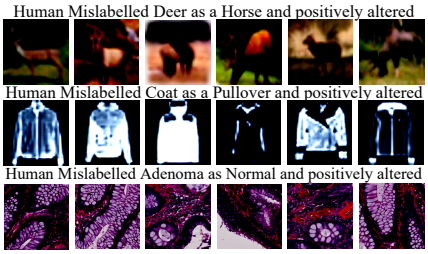

FIGURE 3. Positive alterations made by the L2CU on CIFAR-10N, Fashion-MNIST-H and Chaoyang experiments (top to bottom).

V. RESULTS

Results are reported for unseen test users who were profiled and met the entry condition. Table 1 compares L2CU with competing methods and Table 2 presents the post-alteration accuracy relative to the users' original accuracy, with positive and negative alterations from eq.5 for K selected from silhouette score and Fig.3 with sample positive alterations.

A. COMPARISON WITH RELATED METHODS

Table 1 compares L2CU with competing methods from L2D, L2D with unseen users, and L2C on both simulated (CIFAR-10) and real annotator datasets (CIFAR-10N, CIFAR-10H, Fashion-MNIST-H, Chaoyang). For a fair comparison, Chaoyang results are reported for all users without applying the entry condition. We prioritize comparisons with real annotator datasets and use simulations only when necessary. Some values are missing due to requiring the same users in training and testing sets, not supporting multi-rater settings with sparse labels (where we resort to simulation data), or unavailable source code (where we report values from original work). Unlike literature methods trained with ground truth (except LECODU), our models are trained without it (with consensus) yet still outperform them. Low standard deviation indicate steady improvements across users and datasets.

B. RESULTS OF DATASETS WITH SIMULATED ANNOTATORS

The first row of Table 2 show the number of testing users who improved (I), maintained (M), or did not improved (NI) with L2CU in the CIFAR-10 simulation. The comparison between original and post-alteration accuracy reveals that all 15 testing users improved, with post-alteration accuracy exceeding the original. The last two columns show a large A_+ contrasted with a low A_-, highlighting a high proportion of positive alterations.

C. RESULTS OF DATASETS WITH REAL ANNOTATORS

Table 2 shows that all test users improved their post-alteration accuracy with L2CU by approximately 18%, 5%, 32%, and 15% for CIFAR-10N, CIFAR-10H, Fashion-MNIST-H, and Chaoyang, respectively along with a high positive alteration rate compared to negative alterations. Interestingly, real annotator datasets (except Chaoyang) have low silhouette scores likely because they include many annotators, each introducing subtle noise patterns, making profiles harder to distinguish. Being a real annotator dataset, with fewer annotators, Chaoyang has the highest score. However, L2CU manages to improve users in such challenging conditions.

D. ADAPTING TO THE TEXT DOMAIN

Table 3 evaluates L2CU in the text domain using AgNews [36] dataset following a simulation setup similar to CIFAR-10. All 15 test users improved post-alteration accuracy over original and a high positive alteration proportion, promising that adapting to the text domain is a possibility. See appendix C for experiment setup.

TABLE 3. L2CU performance on Text domain with Agnews

K (Silhouette score)	Test Users	I	M	NI	Original Accuracy	Post. Alt. Accuracy	A_+	A_-
3 (0.44)	15	15	0	0	0.700	0.980	0.975	0.016

E. ROLE OF ANNOTATOR PROFILES

To evaluate the impact of annotator profiles on unseen user performance, we compare results with and without profiles (Tables 2 and 4). The absence of profiles leads to a higher number of users whose performance is not improved (NI), lower post alteration accuracy, and lower positive label alterations across all datasets compared to *with* profiles. This demonstrates that profiles are crucial for improving the performance of unseen users.

To validate the silhouette score's selection of the optimal number of clusters (K), we visualize the profiles identified in the CIFAR-10 simulation experiment with $K = 3$ (Fig. 4, Appendix E). The estimated noise matrices for these profiles closely resemble the ground-truth noise patterns used in the simulation: one profile exhibits noise between airplane and bird, another between horse and deer, and a third between

TABLE 4. Results without annotator profiles

Dataset	I	M	NI	Original Accuracy	Post. Alt. Accuracy	A_+	A_-
CIFAR-10	5	0	10	0.880	0.835	0.815	0.261
CIFAR-10N	69	0	11	0.836	0.918	0.893	0.006
CIFAR-10H	1949	0	73	0.940	0.911	0.882	0.006
F-MNIST-H	166	2	14	0.662	0.854	0.635	0.181
Chaoyang	1	0	1	0.858	0.915	0.704	0.065

truck and automobile. This confirms both the effectiveness of the silhouette score in determining K and the ability of clustering approach to identify distinct user noise patterns.

F. DISTRIBUTION OF JOINT DECISIONS

Table 5 shows decision distributions for unseen test users in real annotator experiments, comparing humans, the base model, and their joint cooperation. Each decision—by the human, base model $f_{\psi_k}(.)$, or cooperation $m_{\theta_k}(.)$—is marked as correct (✓) if it matches the test set target label or incorrect (✗) otherwise.

TABLE 5. Proportion that each combination of Human, AI, or Cooperation is correct ✓or incorrect ✗; Columns sum to 1.

Human	AI $f_{\psi_k}(.)$	Cooperation $m_{\theta_k}(.)$	CIFAR -10N %	CIFAR -10H %	F-MNIST-H %	Chaoyang %
✗	✓	✓	05.15	05.59	04.47	03.35
✓	✗	✓	00.65	02.26	15.05	01.82
✓	✓	✓	93.79	91.35	72.13	92.16
✗	✗	✓	00.05	00.05	04.29	00.13
✗	✓	✗	00.13	00.19	00.33	00.49
✓	✗	✗	00.11	00.39	01.38	01.29
✓	✓	✗	00.00	00.00	00.20	00.00
✗	✗	✗	00.12	00.17	02.17	00.76

According to Table 5, the majority of correct joint decisions and the lowest proportion of incorrect joint decisions occurs when both human and AI predictions are correct, as expected from a cooperation. In addition, joint decisions tend to be correct when at least one party is correct, showing the effectiveness of cooperation. Interestingly, there are cases where the cooperative decision is correct when both individual parties are wrong which we discuss this in Sec.VII.

VI. ABLATION STUDIES

(Ablation 1) Performance vs. Model Components: Table 6 evaluates the importance of Human Label Encoder h_ϕ and Decision Model d_ζ for overall performance by turning each off separately and together with CIFAR-10N. By changing the eq. 3, when h_ϕ is off, $m_{\theta_k}(\mathbf{x}, \hat{\mathbf{y}}) = d_{\zeta_k}(f_{\psi_k}(\mathbf{x}) \oplus \hat{\mathbf{y}})$. When $d_\zeta(.)$ is off, $m_{\theta_k}(\mathbf{x}, \hat{\mathbf{y}}) = f_{\psi_k}(\mathbf{x}) \oplus h_{\phi_k}(\hat{\mathbf{y}})$. When both are off, $m_{\theta_k}(\mathbf{x}, \hat{\mathbf{y}}) = f_{\psi_k}(\mathbf{x}) \oplus \hat{\mathbf{y}}$.

Accuracy becomes lowest with both components off and highest when both are active. Using only one is suboptimal, showing the importance of added components for unseen users. **(Ablation 2) Performance vs. G:** Table 7 extends the CIFAR-10N experiment to study post-alteration accuracy for augmentation times $G \in \{0, 1, 3, 5\}$. Accuracy jumps significantly from $G = 0$ to $G = 1$ showing the effect of noisy label augmentation, with steady gains for $G > 1$. **(Ablation 3) Performance vs. Noise Rate:** Table 8 expands the CIFAR-10 simulation (Section IV) to study asymmetric noise rates (40%–90%). Our approach remains robust, maintaining accuracy above 86% across cases. **(Ablation 4) Evaluating Backbone Models:** We extend the CIFAR-10N experiment with DenseNet-121, ResNet-50, and ViT/B-16 as base model. Tab.9 shows consistent performance across all models, being agnostic to the backbone.

We perform four additional ablations in which we keep the findings here and the results in the Appendix D. **(Ablation 5) Performance vs. K:** Furthering role of clusters in Sec. V-E, we study the effect of having a K, different from the silhouette optimal. Table 10 extends CIFAR-10N experiment to $K \in \{1, 2, 3, 6, 10\}$ showing that post alteration accuracy improves from $K = 1$ (no clusters) to $K = 2$ (optimal for CIFAR-10N), but drops for $K > 2$, likely due to over-adaptation, as higher K reduces training users per profile, making the model less generalizable. **(Ablation 6) Performance vs. λ:** We study the impact of λ in the loss function eq.4 on post-alteration accuracy using CIFAR-10N with ResNet-50, DenseNet-121, and Vit/B-16. Experiments with $\lambda \in \{0, 0.01, 0.1, 1, 10\}$ (Table 11) show that the highest accuracy is centered around $\lambda = 0.1$ across all models. **(Ablation 7) Performance vs. SVM Profiling Error:** To assess the impact of SVM profiling errors, we randomly assigned test users a profile different from the one predicted by the SVM. Table 12 shows the expected surge in not improved (NI) users and the drop in post alteration accuracy. This performance drop contrasts with Table 1, where correct SVM assignments yield higher post-alteration accuracy. **(Ablation 8) Performance vs. Profile Assignment Method:** Hard assignment uses the prediction from the most matching profile model m_{θ_k} at inference, while soft assignment averages predictions from all $m_{\theta_{k\in\{1,\ldots,K\}}}$, weighted by OVA SVM's profile probabilities. Table 13 shows that hard assignment performs better than soft profile assignment across all datasets. All ablation studies adopt the setup in Section IV and use the K selected by silhouette score in Table 2.

VII. DISCUSSION

A. CAN COOPERATION CORRECT JOINT MISTAKES?

Table 5 reveals an interesting phenomenon: the cooperative decision can be correct even when both human and AI are wrong. This happens because the decision model $d_{\zeta_k}(.)$ learns to exploit the joint label noise distribution of the base model and human to correct their combined errors. A necessary condition for this is $P(C|\neg A, \neg B) > 0$, where A, B and C represent events that correct predictions from the AI, human, and human-AI cooperation, respectively. Given that, base model and human can make mistakes and assuming events A and B are independent and independent given C, we trivially obtain: $P(C|\neg A, \neg B) = \frac{P(\neg A, \neg B|C).P(C)}{P(\neg A, \neg B)} = \frac{(1-P(A|C)).(1-P(B|C)).P(C)}{(1-P(A))(1-P(B))} > 0$ because $0 < P(B|C), P(A|C), P(A), P(B), P(C) < 1$.

B. WHEN WOULD COOPERATION REDUCE PERFORMANCE?

Although generally improves joint performance, L2CU may degrade the performance of experts with near-perfect accuracy. We observe this with Chaoyang dataset having one expert with 99.6% original accuracy and 99.3% post alt. accuracy. This aligns with theoretical findings in [5] which demonstrates that an L2C system's improvement is contingent upon the model outperforming its human counterpart. To

TABLE 6. Performance vs. importance of components. Checkmarks indicate (not-)having component--Ablation 1

$h_\phi(.)$	$d_\zeta(.)$	Post Alt. Acc	A_+	A_-
✗	✗	0.705	0.007	0.159
✓	✗	0.774	0.043	0.083
✗	✓	0.861	0.833	0.134
✓	✓	0.989	0.954	0.004

TABLE 7. Performance vs. noisy label augmentation hyper-parameter G --Ablation 2

G	Post Alt. Acc.	A_+	A_-
0	0.615	0.411	0.302
1	0.980	0.953	0.004
3	0.983	0.954	0.004
5	0.989	0.952	0.004

TABLE 8. Performance vs. noise rate --Ablation 3

Noise Rate	Post Alt. Acc.	A_+	A_-
40%	0.992	0.973	0.001
60%	0.968	0.953	0.088
80%	0.879	0.944	0.134
90%	0.868	0.875	0.195

TABLE 9. Performance vs. different backbones as base model $f_{\psi_k}(.)$ --Ablation 4

Backbone Model	Post Alt. Acc.	A_+	A_-
ResNet-50	0.968	0.862	0.013
DenseNet-121	0.969	0.854	0.011
Vit/B-16	0.989	0.954	0.004

mitigate degrades, we introduced an entry condition in our profiling process (Sec. III-C) that compares the base model accuracy on $\mathcal{V}$, to that of the test user.

C. FUTURE DIRECTIONS

First, extending L2CU to a distribution-agnostic user setting would be a promising direction that broadens its application. Second, modeling temporal dynamics would enable complementing the changing user behavior over time. Third, using meta-learning to adapt AI cooperative models to profiles would lower the computational overhead. Finally, developing a few-shot user profiling would reduce the annotation requirements, especially for datasets with many classes.

VIII. CONCLUSION

This paper introduced L2CU, a novel learning to complement framework that enables complementing unseen users in human-AI cooperative classification. Extensive evaluations across datasets (CIFAR-10N, CIFAR-10H, Fashion-MNIST-H, Chaoyang, and AgNews) demonstrate L2CU's leading performance without needing ground truth. Furthermore, the proposed label augmentation method tackles data sparsity while preserving annotator bias, and the alteration rate metric offers insights into the model's impact on human labels. With the model-agnostic design and ability to leverage noisy, sparse multi-user data without access to ground truth, L2CU offers a significant step toward human-AI cooperative classification systems.

APPENDIX A
CONSENSUS LABEL ESTIMATION

Many multi-rater input datasets lack ground truth labels. To address this, L2CU is built to function effectively without relying on them. During training, we use Crowdlab [21] to estimate a consensus label $\bar{\mathbf{y}}_i$, which approximates the true clean label $\mathbf{y}_i$. [21] works in two steps. In the first step, it estimates a consensus by majority vote $\bar{\mathbf{y}}_i'$ per training sample. In the second step, it trains a classifier using the initial consensus and obtains predicted class probabilities for each training example. Thereafter, these predicted probabilities along with the original annotations from the raters are used to estimate a better consensus, creating the following ensemble,

$$\bar{\mathbf{y}}_i = \mathbf{w}_\gamma \times f_\gamma(\mathbf{x}_i) + \mathbf{w}_1 \times \tilde{\mathbf{y}}_{i,1} + \ldots + \mathbf{w}_{|\mathcal{A}|} \times \tilde{\mathbf{y}}_{i,|\mathcal{A}|}, \quad (6)$$

where $f_\gamma : \mathcal{X} \to \Delta^{C-1}$ is a classifier trained with the majority vote label $\bar{\mathbf{y}}_i'$ to output a categorical distribution for C classes, and the weights $\mathbf{w}_\gamma, \mathbf{w}_1, \ldots, \mathbf{w}_{|\mathcal{A}|}$ are assigned according to an estimate of how trustworthy the model is, compared to each individual annotator. The outcome of Crowdlab is a consensus labeled training set denoted by $\bar{\mathcal{D}} = \{(\mathbf{x}_i, \bar{\mathbf{y}}_i)\}_{i=1}^N$. Note that the consensus label is necessary only when the clean label $\mathbf{y}_i$ is latent. If such clean label is observed, then Crowdlab is no longer needed, and L2CU can be trained with $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$.

APPENDIX B
DECIDING THE OPTIMAL NUMBER OF PROFILES

We determine the optimal number of profiles K with the silhouette score defined by,

$$S_k = \frac{1}{|A|} \sum_{j \in \mathcal{A}} \frac{b(\mathbf{s}_j) - a(\mathbf{s}_j)}{\max\{a(\mathbf{s}_j), b(\mathbf{s}_j)\}}, \quad (7)$$

where $a(\mathbf{s}_j)$ denotes the sample's intra-profile distance (i.e., the average L2 distance to all other points in the same profile), $b(\mathbf{s}_j)$ represents the inter-profile distance (i.e., the lowest average L2 distance to all points in any other profile). The mean silhouette score for K profiles is defined by $S(K) = \frac{1}{K}\sum_{k=1}^K S_k$. The optimal number of profiles for the dataset is identified by the K that yields the highest silhouette score.

APPENDIX C
EXPERIMENTAL SETUP

A. SETUP FOR DATASETS WITH REAL ANNOTATORS

When training with CIFAR-10N, we present two experiments. For the first experiment, the labels from 747 annotators form $\tilde{\mathcal{D}}$. Out of them, 159 were identified for having annotated at least 20 images per class, and they were split in half, taking 79 as training users and 80 as testing users. The training users' labels are used to build the K profiles and train the OVA SVM classifier, where K is automatically chosen based on the silhouette score in eq. 7. During testing, a testing user's noisy-label transition matrix is estimated using the test annotator's labels and consensus labels. This matrix is used to simulate noisy annotations from that testing user. Therefore, 80 noisy test sets are produced, with each representing the biases that user possesses. The model for each profile k, denoted by $m_{\theta_k}(.)$, uses ViT-Base-16 [37] as the $f_{\psi_k}(.)$.

For the second CIFAR-10N experiment, we use CIFAR-10H as the testing set, where the labels from testing users were used without any modification for user profiling. The same labels were used to estimate a noise transition matrix and simulate their own test set. For all 2571 users, their own

TABLE 10. Performance vs. number of profiles K on CIFAR-10N --Ablation 5

K	Post Alt. Accuracy.	A_+	A_-
K=1	0.918	0.893	0.006
K=2	0.989	0.954	0.004
K=3	0.988	0.954	0.004
K=6	0.986	0.944	0.004
K=10	0.972	0.914	0.004

TABLE 11. Performance vs. λ of the loss in eq.4 (with CIFAR-10N) --Ablation 6

Backbone model	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$
ResNet-50	0.929	0.944	0.968	0.939	0.929
DenseNet-121	0.936	0.950	0.969	0.937	0.931
ViT-B/16	0.982	0.982	0.989	0.976	0.969

TABLE 12. Performance vs. SVM profiling error. --Ablation 7

Dataset	Not Improved Users (NI)	Post Alt. Accuracy
CIFAR-10	11	0.82
CIFAR-10N	29	0.92
CIFAR-10H	181	0.90
F-MNIST-H	71	0.84
Chaoyang	All	0.89

TABLE 13. Performance vs. profile assignment method. --Ablation 8

Profile Assignment Method	CIFAR -10N	CIFAR -10H	F-MNIST-H	Chaoyang
Soft	0.982	0.989	0.866	0.945
Hard	0.989	0.993	0.878	0.988

test test was simulated that possess own biases. The models trained for CIFAR-10N were used for this experiment.

For the Fashion-MNIST-H experiment, the labels from all 885 annotators are taken to form the $\tilde{\mathcal{D}}$. Then, 366 out of 885 users are chosen since they have annotated at least 20 images per class and are split in half to have 183 users for training and 183 for testing. The training users' labels are used to build the K profiles and train the OVA SVM classifier, where K is automatically chosen based on the silhouette score in eq.7. During testing, the testing user's noisy-label transition matrix is estimated using the annotator's labels and consensus labels. This matrix is used to simulate test annotations from that testing user. Therefore, 183 noisy testing sets are produced, with each representing the biases that each user possesses. The model for each profile k, represented by $m_{\theta_k}(.)$ uses DenseNet-121 [38] for $f_{\psi_k}(.)$.

Chaoyang has three annotators per image, which form the $\tilde{\mathcal{D}}$. Training users are used to make K profiles, and train an OVA SVM, where K is automatically chosen based on the silhouette score in eq.7. For each profile k, a model $m_{\theta_k}(.)$ is trained with a ViT-Large-16 as the backbone for $f_{\psi_k}(.)$. During testing, user's noisy-label transition matrix is estimated using the annotator's labels and consensus labels. This matrix is used to simulate noisy test annotations from that user, resulting three noisy test sets.

Our experiment with CIFAR-10N and CIFAR-10H, with human labels for CIFAR-10's training and testing sets respectively, offer a more realistic setup, better reflecting real-world conditions. But, while our method preserves annotators' noisy label patterns, it's important to note that Fashion-MNIST-H and Chaoyang test sets are simulated and might not completely mimic real annotator inputs.

In our CIFAR experiments, we adopted the data augmentation policy introduced by [33]. Also, for Fashion-MNIST, alongside random horizontal and vertical flips, we integrated auto augmentations as proposed by [34]. For the Chaoyang dataset, data augmentation was limited to random resized crops of dimensions 224×224. We rely on pre-trained models for f_{ψ_k} because of their robustness to noisy labels [39] (e.g., ViT models, ResNet-18 and DenseNet-121 models are pre-trained on ImageNet-1K.) Adam optimizer was employed for training $f_{\psi_k}(.)$ with consensus $\bar{\mathcal{D}}$, where NAdam was used for training $m_{\theta_k}(.)$ on $\hat{\mathcal{D}}$, each utilizing their respective default learning rates. Implementations were done in PyTorch and executed on single GeForce RTX 4090 GPU.

B. SETUP FOR EXPERIMENT IN TEXT DOMAIN

This experiment evaluates L2CU in the text domain using AgNews [36] following a simulation setup similar to CIFAR-10. AgNews is a text classification dataset with 120,000 training, 200 validation and 7,400 testing class-balanced news articles categorized into 4 classes. We perform pairwise label flipping on two out of four classes, where 60% of samples are flipped to the incorrect class while the other two classes remained clean. We simulate three profiles of users, one that flips between classes business↔science/technology, another profile that flips world↔sports, and the third profile that flips sports↔business. Five training and five testing users are simulated for each profile producing a total of 15 unique training and testing users. The title and description were concatenated and truncated to a maximum of 64 tokens, Bert-Tokenizer was used to tokenize and Bert-Base-Uncased [40] model is used as base model $f_{\psi_k}(.)$ when training. All 15 test users showed improved post-alteration accuracy over original and a high positive alteration proportion, indicating while further future experiments are needed, adapting to the text domain is a possibility.

APPENDIX D
ADDITIONAL ABLATION STUDIES

This section provides the results for four additional ablation studies: (1) post alteration accuracy vs. having a different K than the optimal from the silhouette score (Ablation 5 in Table 10), (2) post alteration accuracy vs. λ in loss function eq.4 (Ablation 6 in Table 11), (3) post alteration accuracy vs. SVM profiling errors (Ablation 7 in Table 12), and (4) post alteration accuracy vs. profile assignment method being soft or hard (Ablation 8 in Table 13) ; that we discussed in Sec. VI

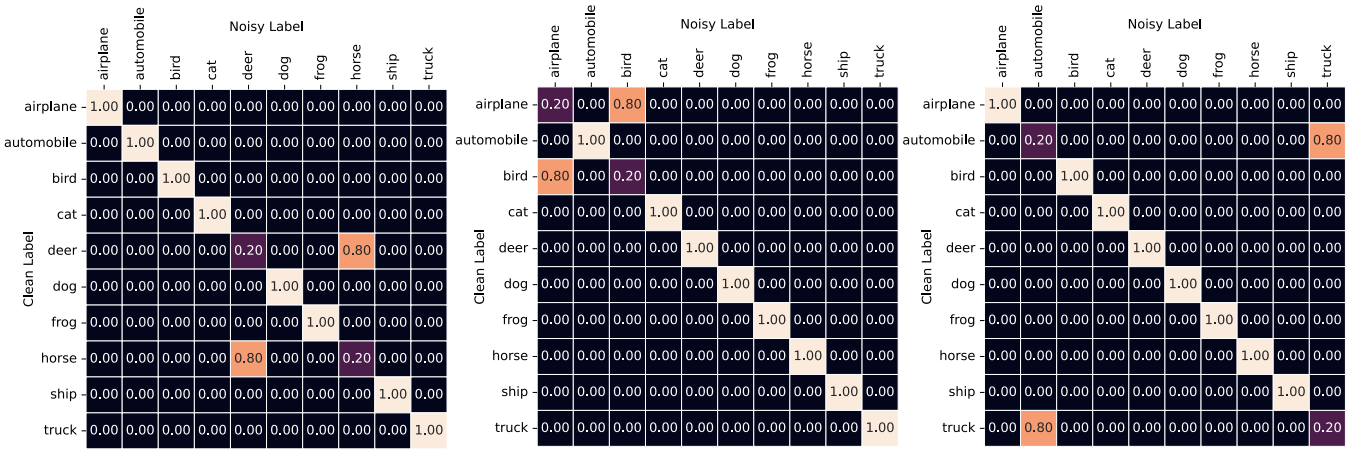

FIGURE 4. **Estimated noise matrices for identified annotator profiles from CIFAR-10 simulation experiment.**

in the main paper.

APPENDIX E
VISUALIZING NOISE PROFILES

Figure 4 shows the estimated noise matrices of identified annotator profiles from the CIFAR-10 simulation with $K = 3$, selected via the silhouette score. Notably, these matrices closely resembles the simulated noise patterns used to create 15 users—flipping between airplane↔bird, horse↔deer, and truck↔automobile. This validates the ability of the clustering to capture distinct noise patterns and confirms the effectiveness of the silhouette score in selecting an optimal K for accurate user profiling.

APPENDIX F
EXTENDED RELATED WORK

While restating that the goal of L2CU is improving human-AI joint decision-making with unseen users, we review related yet distinct research areas due to the interdisciplinary nature of the problem.

Learning from Noisy-labels (LNL) aims to design algorithms that are robust to the presence of noisy training labels. Recent advancements include DivideMix [41] with its semi-supervised approach, ELR [42] exploring early learning phenomena with a regularised loss, CausalNL [43] proposing a generative model for instance-dependent label-noise learning, C2D [44] tackling the warm-up obstacle, UNICON [22] with a unified supervised and unsupervised learning, and [45] with graphical modeling for noise rate estimation to handle noisy labels effectively. In contrast, L2CU does not aim to infer clean labels, but to improve human-AI joint decision-making by adapting to annotator-specific biases—shifting the goal from label denoising to cooperation.

Multi-rater Learning (MRL) leverages noisy labels from multiple annotators per sample, often to estimate consensus or to mitigate the identifiability problem under certain conditions [46]. Key developments include, [47], with expectation maximization algorithm to estimate ground truth, MRNet [48], which addresses multi-rater disagreement, [21], estimating consensus being model-agnostic in design, [49] adopting a Mixture of Experts architecture for learning from multiple noisy sources, and [50] addressing the sparse crowd annotations. We employ [21] to estimate consensus in L2CU for its model-agnostic nature and good performance, although it is an open choice in L2CU framework. Although consensus estimation is a preliminary step, L2CU's core contribution lies in complementing annotator-specific behavior—an objective different from MRL.

Learning with crowds (LWC) aims to train models with sparse labels provided by multiple annotators. [51] and [52] aim to aggregate diverse annotator signals while addressing label reliability. Sel-CL [53] tackles noisy supervision by leveraging contrastive learning to better separate clean and corrupted labels. AIDTM [54] learns annotator-specific noise patterns, while CCC [55] handles settings with limited labels per annotator through parameter-efficient modeling. However, unlike LWC, that aims to denoise annotations, L2CU preserves individual annotator signals to train profile-specific cooperative AI models that can adapt to and complement unseen users.

Despite improvements from LNL, MRL and LWC, an accuracy gap persists compared to training with clean labels. This has led to our human-AI joint decision-making paradigm, which incorporates inputs from both humans and AI to make decisions. Unlike objectives of recommendation systems, LWC, LNL or MRL—that focus on preference modeling, inferring latent ground truth, being robust to label noise, or consensus estimation—our goal is to improve human-AI joint decision-making by complementing unseen users—the fundamental distinction of L2CU.

REFERENCES

[1] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: Improving fairness and accuracy by learning to defer," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 6150–6160.

[2] H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag, "Who should predict? exact algorithms for learning to defer to

humans," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, F. Ruiz, J. Dy, and J.-W. van de Meent, Eds., vol. 206. PMLR, 25–27 Apr 2023, pp. 10 520–10 545. [Online]. Available: https://proceedings.mlr.press/v206/mozannar23a.html

[3] J. V. Alves, D. Leitão, S. Jesus, M. O. Sampaio, P. Saleiro, M. A. Figueiredo, and P. Bizarro, "Fifar: A fraud detection dataset for learning to defer," *arXiv preprint arXiv:2312.13218*, 2023.

[4] B. Wilder, E. Horvitz, and E. Kamar, "Learning to complement humans," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, 2021.

[5] M. Steyvers, H. Tejeda, G. Kerrigan, and P. Smyth, "Bayesian modeling of human-AI complementarity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 11, p. e2111547119, Mar. 2022.

[6] J. Strong, P. Saha, Y. Ibrahim, C. Ouyang, and A. Noble, "Expert-agnostic learning to defer," *arXiv preprint arXiv:2502.10533*, 2025.

[7] D. Tailor, A. Patra, R. Verma, P. Manggala, and E. Nalisnick, "Learning to defer to a population: A meta-learning approach," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 3475–3483.

[8] B. Strauch, "Ironies of automation: Still unresolved after all these years," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 419–433, 2018.

[9] M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan, "The algorithmic automation problem: Prediction, triage, and human effort," *arXiv preprint arXiv:1903.12220*, 2019.

[10] N. Okati, A. De, and M. Rodriguez, "Differentiable learning under triage," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9140–9151, 2021.

[11] H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. Iii and A. Singh, Eds., vol. 119. PMLR, 2020, pp. 7076–7087.

[12] R. Verma and E. Nalisnick, "Calibrated learning to defer with one-vs-all classifiers," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 22 184–22 202. [Online]. Available: https://proceedings.mlr.press/v162/verma22c.html

[13] C. C. Nguyen, T.-T. Do, and G. Carneiro, "Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=zl0HLZOJC9

[14] Z. Zhang, C. Nguyen, K. Wells, T.-T. Do, and G. Carneiro, "Learning to complement with multiple humans," *arXiv preprint arXiv:2311.13172*, 2023.

[15] Z. Zhang, W. Ai, K. Wells, D. Rosewarne, T.-T. Do, and G. Carneiro, "Learning to complement and to defer to multiple users," in *European Conference on Computer Vision*. Springer, 2024, pp. 144–162.

[16] R. Kocielnik, S. Amershi, and P. N. Bennett, "Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, no. Paper 411. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–14.

[17] D. Wang, L. Wang, Z. Zhang, D. Wang, H. Zhu, Y. Gao, X. Fan, and F. Tian, ""brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, no. Article 697. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–18.

[18] G. Kerrigan, P. Smyth, and M. Steyvers, "Combining human predictions with model probabilities via confusion matrices and calibration," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4421–4434, 2021.

[19] K. Vodrahalli, T. Gerstenberg, and J. Y. Zou, "Uncalibrated models can improve human-ai collaboration," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4004–4016, 2022.

[20] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[21] H. W. Goh, U. Tkachenko, and J. Mueller, "Crowdlab: Supervised learning to infer consensus labels and quality scores for data with multiple annotators," *NeurIPS 2022 Human in the Loop Learning Workshop*, 2022.

[22] N. Karim, M. N. Rizve, N. Rahnavard, A. Mian, and M. Shah, "Unicon: Combating label noise through uniform selection and contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9676–9686.

[23] J. Xu, J. Han, K. Xiong, and F. Nie, "Robust and sparse fuzzy k-means clustering," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 2224–2230.

[24] V. K. Dehariya, S. K. Shrivastava, and R. Jain, "Clustering of image data set using k-means and fuzzy k-means algorithms," in *2010 International conference on computational intelligence and communication networks*. IEEE, 2010, pp. 386–391.

[25] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, Ontario, Tech. Rep., 2009.

[27] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu, "Learning with noisy labels revisited: A study using real-world human annotations," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=TBWA6PLJZQm

[28] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[29] T. Ishida, I. Yamane, N. Charoenphakdee, G. Niu, and M. Sugiyama, "Is the performance of my deep network too good to be true? a direct approach to estimating the bayes error in binary classification," in *ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=FZdJQgy05rz

[30] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[31] C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin, "Hard sample aware noise robust learning for histopathology image classification," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 881–894, 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[34] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[36] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *NIPS*, 2015.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[39] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *International conference on machine learning*. PMLR, 2020, pp. 4804–4815.

[40] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[41] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJgExaVtwr

[42] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," *Advances in neural information processing systems*, vol. 33, pp. 20 331–20 342, 2020.
[43] Y. Yao, T. Liu, M. Gong, B. Han, G. Niu, and K. Zhang, "Instance-dependent label-noise learning under a structural causal model," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4409–4420, 2021.
[44] E. Zheltonozhskii, C. Baskin, A. Mendelson, A. M. Bronstein, and O. Litany, "Contrast to divide: Self-supervised pre-training for learning with noisy labels," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1657–1667.
[45] A. Garg, C. Nguyen, R. Felix, T.-T. Do, and G. Carneiro, "Instance-dependent noisy-label learning with graphical model based noise-rate estimation," in *European Conference on Computer Vision*. Springer, 2024, pp. 372–389.
[46] Y. Liu, H. Cheng, and K. Zhang, "Identifiability of label noise transition matrix," in *International Conference on Machine Learning*. PMLR, 2023, pp. 21 475–21 496.
[47] V. B. Sinha, S. Rao, and V. N. Balasubramanian, "Fast dawid-skene: A fast vote aggregation scheme for sentiment classification," 2018.
[48] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.
[49] Y. Zhao, G. Zheng, S. Mukherjee, R. McCann, and A. Awadallah, "Admoe: Anomaly detection with mixture-of-experts from noisy labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 4937–4945.
[50] H. Zhang, S. Li, D. Zeng, C. Yan, and S. Ge, "Coupled confusion correction: Learning from crowds with sparse annotations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 16 732–16 740.
[51] H. Wei, R. Xie, L. Feng, B. Han, and B. An, "Deep learning from multiple noisy annotators as a union," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 552–10 562, 2023.
[52] S. Li, T. Liu, J. Tan, D. Zeng, and S. Ge, "Trustable co-label learning from multiple noisy annotators," *IEEE Transactions on Multimedia*, vol. 25, pp. 1045–1057, 2021.
[53] S. Li, X. Xia, S. Ge, and T. Liu, "Selective-supervised contrastive learning with noisy labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 316–325.
[54] S. Li, X. Xia, J. Deng, S. Ge, and T. Liu, "Transferring annotator-and instance-dependent transition matrix for learning from crowds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 11, pp. 7377–7391, 2024.
[55] H. Zhang, S. Li, D. Zeng, C. Yan, and S. Ge, "Coupled confusion correction: Learning from crowds with sparse annotations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 2024, pp. 16 732–16 740.

DILEEPA PITAWELA received the B.Sc. degree in Information Technology from the University of Moratuwa, Sri Lanka, in 2018, and the M.S. degree in Data Science from the University of Adelaide, Australia, in 2022. He is currently pursuing a Ph.D. degree at the University of Adelaide, Australia. His research interests include computer vision and human–AI collaboration.

GUSTAVO CARNEIRO received the PhD degree in computer science from the University of Toronto, in 2004. He is currently a full professor of AI and Machine Learning with the Centre for Vision, Speech and Signal Processing (CVSSP), The University of Surrey, U.K. Previously, he was a professor with the University of Adelaide, an ARC future fellow, and the director of Medical Machine Learning with the Australian Institute of Machine Learning (AIML). He joined the University of Adelaide as a senior lecturer, in 2011 and became an associate professor, in 2015 and a professor, in 2019. In 2014 and 2019, he joined the Technical University of Munich as a visiting professor and a Humboldt fellow. From 2008 to 2011 he was a Marie Curie IIF fellow and a visiting assistant professor with the Instituto Superior Tecnico (Lisbon, Portugal) within the Carnegie Mellon University-Portugal program (CMU-Portugal). From 2006 to 2008, he was a research scientist with Siemens Corporate Research in Princeton. In 2005, he was a postdoctoral fellow with the University of British Columbia and with the University of California San Diego. His research interests lie in the fields of computer vision, medical image analysis, and machine learning.

HSIANG-TING (TIM) CHEN received the Ph.D. degree in computer science from National Tsing Hua University, Taiwan, in 2013. Afterward, he held research appointments at the University of Tokyo and the Hasso Plattner Institute, where his work advanced digital sketching, 3D modeling, and personal fabrication, which led to multiple patents. He joined the University of Technology Sydney in 2016 as a Lecturer, expanding his research into virtual/augmented reality and brain–computer interfaces within the Centre for Artificial Intelligence. In 2020, he moved to the University of Adelaide, where he leads the Human–Computer Interaction Laboratory. He is also a member of the Australian Institute for Machine Learning and the Robinson Research Institute. Dr. Chen is a Chief Investigator of the 2023 Eureka Prize–winning IMAGENDO Project, leading AI development for sonographer guidance in endometriosis diagnosis. He also serves as the University of Adelaide's research lead for the Augmenting Ability CRC, advancing adaptive AI for decision support in aged and disability care. His current research focuses on leveraging emerging technologies to create interactive AI systems that collaborate with knowledge workers, with a particular emphasis on healthcare.