ECCV 2018 8-14 September Munich, Germany

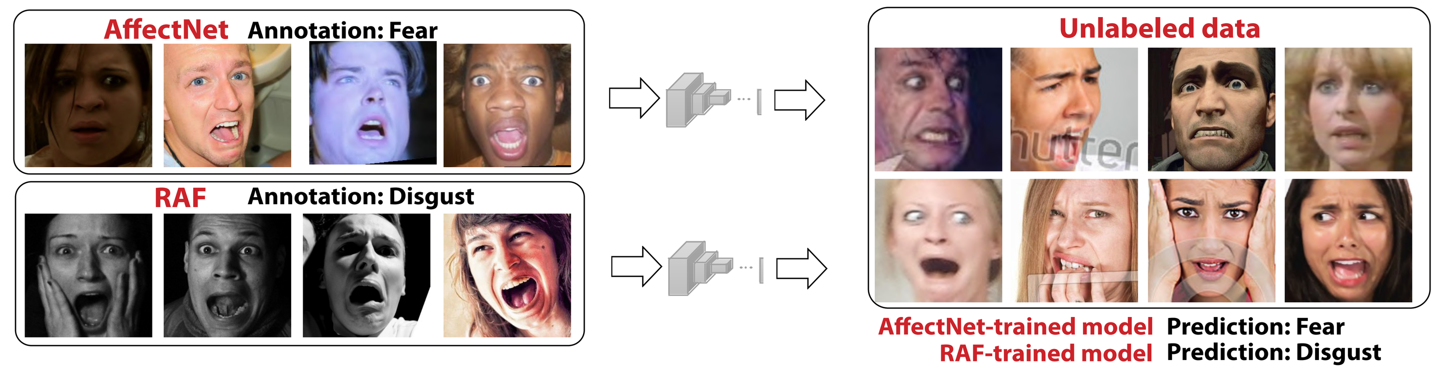
Facial Expression Recognition with Inconsistently Annotated Datasets

Jiabei Zeng, Shiguang Shan, Xilin Chen

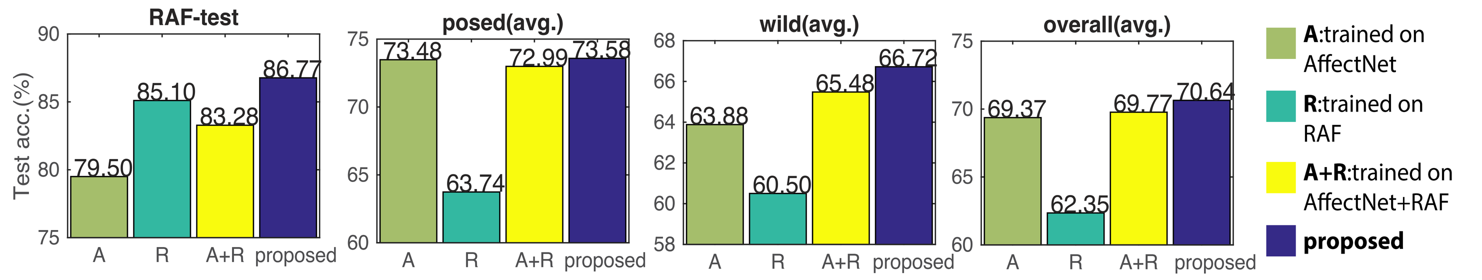
中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Problem

- ❑ Emotion recognition supervised by **more than one** mannally annotated datasets
- ❑ **Challenges**
  - ■ **Errors and bias** of human annotations exist among different facial expression datasets.
    - It is subjective to classify faces into several emotional categories.
    - Human's understanding of facial expressions varies with different cultures, living environments, and their experiences.
  - ■ **Annotation bias** of training datasets ⟶ **recognition bias** of trained systems



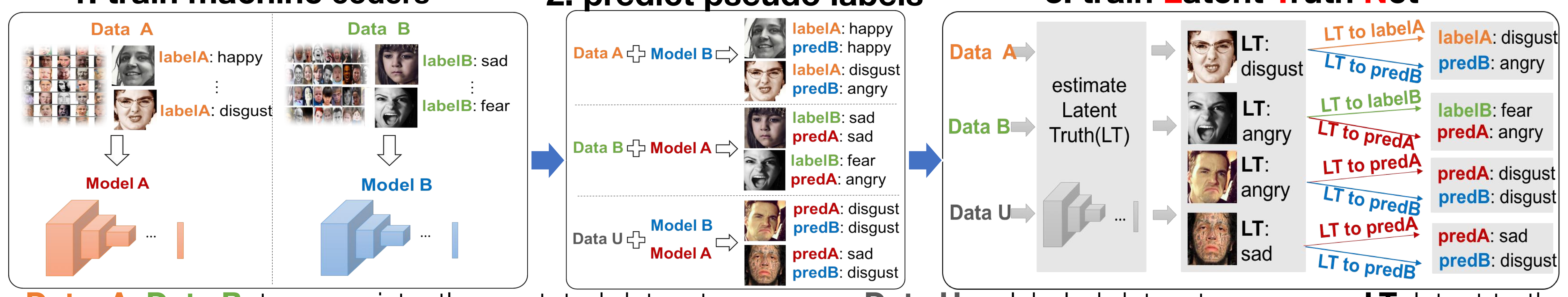
  - ■ **More data** by merging multiple training datasets **≠ better performance** of the trained system



- ❑ **How to learn from multiple datasets with annotation biases?**
  - ■ Learn from noisy labels
    - They leverage a small set of clean data or assume constrains or distributions of the noise.
    - Each sample has one annotation. They neglect the noise pattern from multiple annotations.
  - ■ Learn from crowd sourcing
    - They estimate the true labels of the noisily labeled data by multiple annotators.
    - They focus on estimating the true labels of annotated samples rather than train a model to predict unlabeled data.

Inconsistent Pseudo Annotations to Latent Truth(IPA2LT) framework

- ❑ Learn from multiple coders: human coders, trained models as machine coders
- ❑ Unlabelled data: bridge between the different datasets by sharing the same machine coders with them
- ❑ Annotation bias are modelled as the probability transitions between the latent truth and the (human or machine) coders' annotations in **Latent Truth Net**



Data A, **Data B**: two consistently annotated datasets **Data U**: unlabeled dataset **LT**: latent truth
labelA, **labelB**: human labels for data A and B **predA**, **predB**: predictions by model A and B

Latent Truth Net (LTNet)

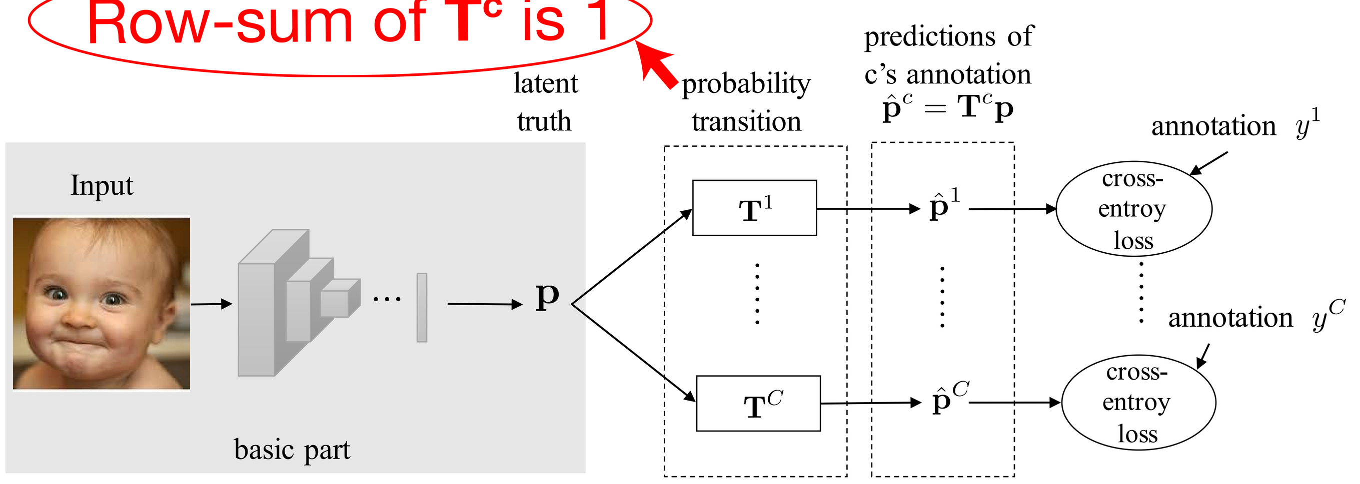
- ❑ Goal: Learn from samples with multiple inconsistent annotations
- ❑ Definition of inconsistent annotations
  - ■ Data: $\mathcal{X} = \{\mathbf{x}_i, \dots, \mathbf{x}_N\}$
  - ■ Each sample $\mathbf{x}_n$ is labelled by C coders with annotations $y_n^1, \dots, y_n^C$
  - ■ Probability distribution of coder i labelling $\mathbf{x}_n : P(y_n^i|\mathbf{x}_n)$
  - ■ Inconsistent annotation assumes
    $P(y_n^i|\mathbf{x}_n) \neq P(y_n^j|\mathbf{x}_n), \forall \mathbf{x}_n \in \mathcal{X}, i \neq j$
- ❑ Formulation of LTNet
  - ■ LTNet learns the latent truth $\mathbf{p}$
  - ■ The bias of coder c is represented by a probability transition matrix $\mathbf{T}^c$. Then, the predicted distribution of c's annotation is
    $$\hat{\mathbf{p}}^c = \mathbf{T}^c\mathbf{p}$$
  - ■ LTNet aims to find the optimal network parameters $\boldsymbol{\Theta}$ and $\mathbf{T}^1, \dots, \mathbf{T}^C$

$$\min_{\Theta, \{\mathbf{T}^1, \cdots, \mathbf{T}^C\}} -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{L}\mathbf{1}(y_n^c = k)\log(\hat{p}_n^c(k))$$

$$s.t. \quad \sum_{j}^{L}\tau_{ij}^c = 1, \forall i = 1, \dots, L$$

Row-sum of $\mathbf{T}^c$ is 1

- ❑ Architechture of LTNet
  - ■ A combination of neural network and Dawid&Skene's[1] truth estimation technique
  - ■ End-to-end trainable



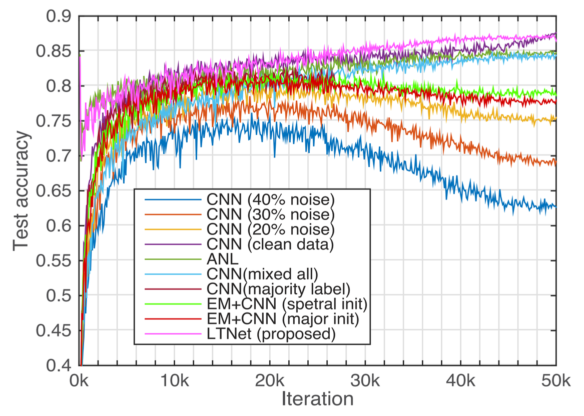
[1] Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Applied statistics pp. 20-28 (1979)

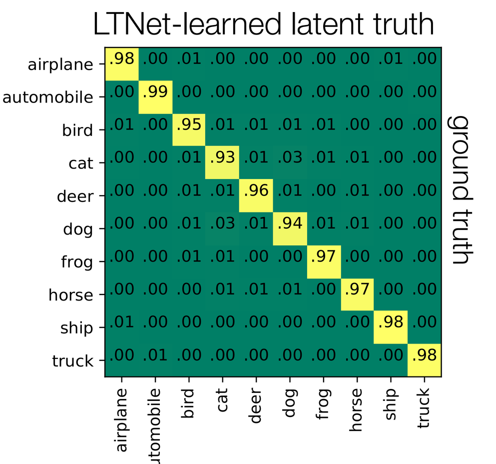
Experiments

- ❑ Synthetic data Code is available at: https://github.com/dualplus/LTNet
  - ■ Randomly revise 20%,30%, or 40% labels in the training set of CIFAR-10.
  - ■ Evaluate the methods on the clean test set.

Table 1: Test accuracy on CIFAR-10 with noisy labels

Training data	Methods	Test Acc.(%)
clean	*basic CNN*	*87.43*
40% noise (A)	basic CNN	62.50
30% noise (B)	basic CNN	68.70
20% noise (C)	basic CNN	75.15
mixture of ABC	basic CNN	84.04
mixture of ABC	AIR(ICLR'16)	76.37
mixture of ABC	NAL(ICLR'17)(re-implementation)	84.41
majority ratings of ABC	basic CNN	82.95
inconsistent annotations	EM+CNN(majority init.)	77.65
inconsistent annotations	EM+CNN(spectral init.)	78.82
inconsistent annotations	LTNet (proposed)	**87.23**



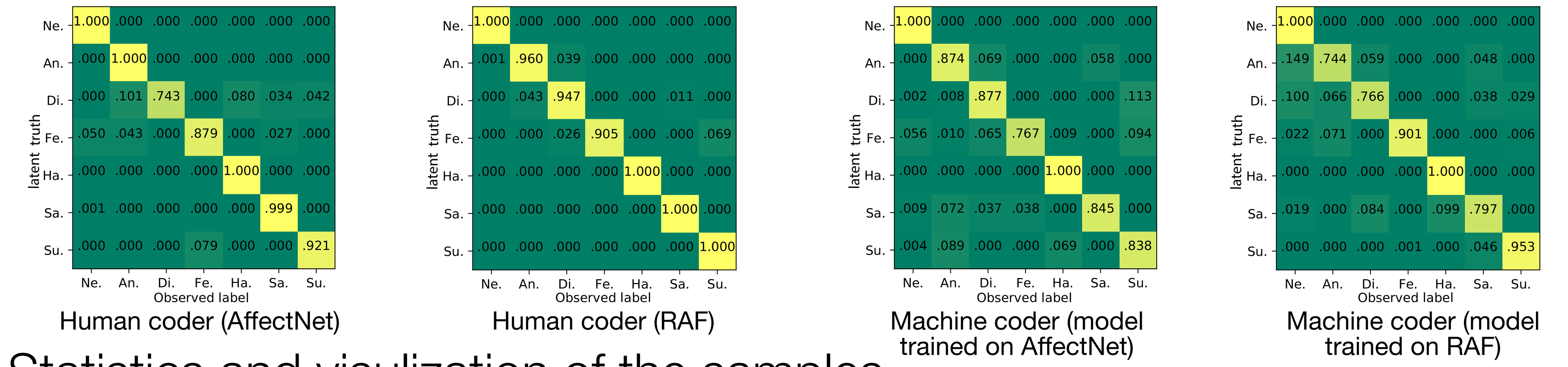


- ❑ Facial expression recognition
  - ■ Training data: AffectNet(training), RAF(training), unlabelled data(~1,200,000)

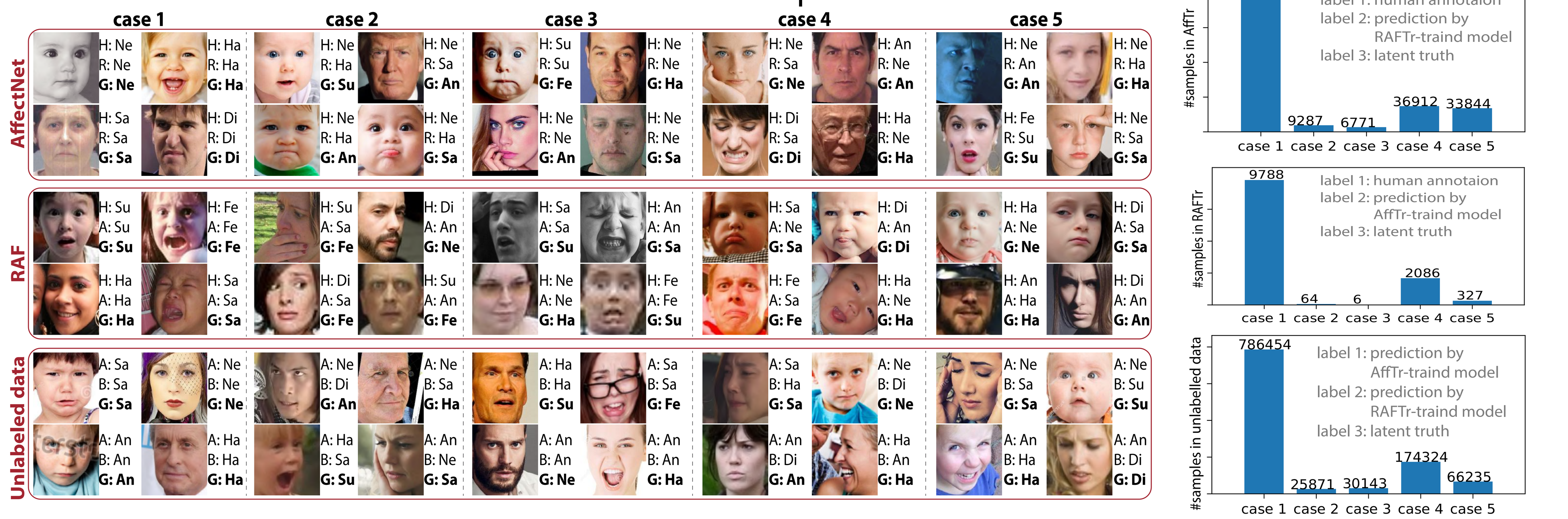
Table 2: Test accuracy on facial expression recognition datasets. (**Bold**: best, Underline: 2nd best)

Methods \ Test sets	in-the-wild RAF (te.)	in-the-wild AffectNet (val.)	in-the-wild SFEW (tr+val)	Posed CK+	Posed CFEE	Posed MMI	Posed Oulu-CASIA	average wild	average posed	average overall
AffTr (base)	79.50	56.51	55.64	91.04	**76.09**	<u>65.32</u>	**61.49**	63.88	<u>73.48</u>	69.37
RAFTr (base)	85.10	44.66	51.75	79.87	64.41	58.17	52.50	60.50	63.74	62.35
AffTr+RAFTr (base)	83.28	<u>56.57</u>	56.58	**92.45**	**76.09**	62.90	60.50	65.48	72.99	69.77
E2E-FC	23.99	24.00	22.33	51.73	26.52	22.25	31.28	23.44	32.95	28.87
AIR(ICLR'16)	67.37	54.23	49.88	43.87	64.47	59.64	47.03	57.16	53.75	55.21
NAL(ICLR'17)	84.22	55.97	<u>58.13</u>	91.20	75.84	64.71	61.00	<u>66.11</u>	73.19	<u>70.15</u>
IPA2LT(EM+CNN)	<u>85.30</u>	**57.31**	54.94	86.64	72.48	63.11	59.95	65.85	70.54	68.53
IPA2LT(LTNet)	**86.77**	55.11	**58.29**	<u>91.67</u>	<u>76.02</u>	**65.61**	<u>61.02</u>	**66.72**	**73.58**	**70.64**

  - ■ LTNet-learned transition matrix T for 4 coders



  - ■ Statistics and visulization of the samples



case 1: label 1=label 2=label 3 **case 2**: label 1≠label 2≠label 3 **case 3**: label 3≠label 2=label 1 **case 4**: label 3=label 1≠label 2 **case 5**: label 3=label 2≠label 1