Implementation Details

Table 1 shows all parameters used for each dataset. All code, datasets, teacher models and experiment settings are available in the file "ska_supplement_material.zip". We also publish all of the code anonymously at https://github.com/hiddenforreview/SKA.

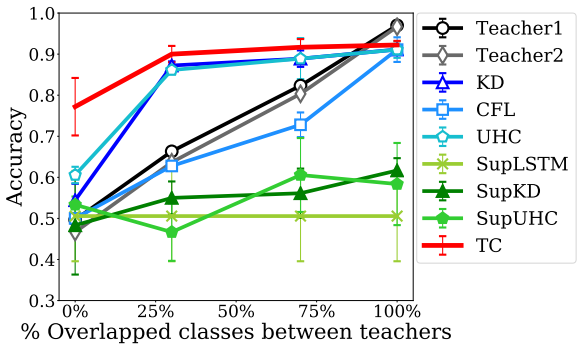
Parameters	SYN	PED	HAR	ELEC
Teacher Trust Learner				
Layers	1	3	2	2
Hidden units	8	8	8	8
Learning rate	0.01	0.001	0.001	0.001
Batch size	8	16	32	64
Epochs	1,000	3,000	1,000	2,000
Student Network				
Layers	1	3	2	2
Hidden units	8	8	8	8
Learning rate	0.01	0.001	0.001	0.001
Batch size	8	16	32	64
Epochs	500	200	200	400

Table 1: Parameters used for the Teacher Trust Learner and the Student Network on each dataset.

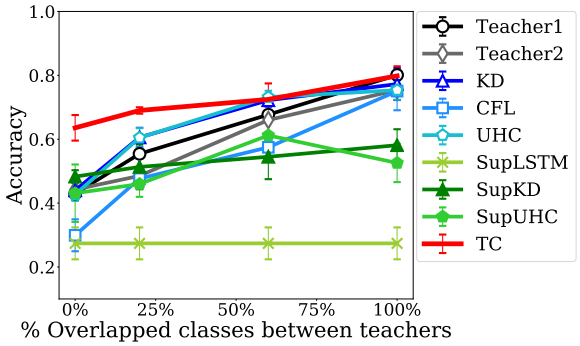
Additional Results

Amalgamating disparate teachers.

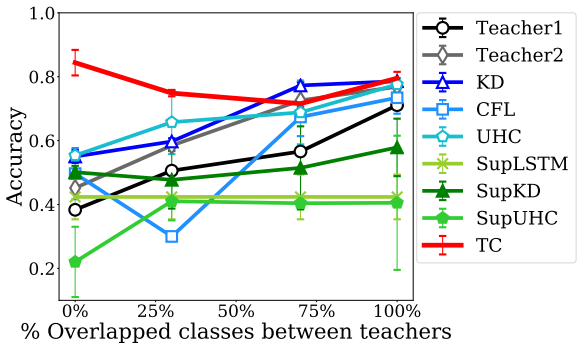
We observe the impacts of amalgamating teachers that are trained independently on their own tasks and therefore provide unrelated outputs, expanding on the results presented in Figure 4 in our full paper. These experiments vary the *proportion of overlapping classes* between the teachers and we assume that the smaller the overlap between predicted classes, the more *disparate* are the teachers because there is less information shared between the teachers tasks. The results of all datasets on the full range class set overlaps (0% to 100%) are shown in Figure 1, assuming that all methods have access to annotations for only 2% of the training data. Across all datasets, we first observe that the supervised methods (the green lines) fair quite poorly compared to the KA methods (blue lines and our approach, the red line). As shown in Table 2 in the full paper, this is because the low level of annotation here (2%) is not enough to train the robust models from scratch. Second, we observe that as overlap between classes increases, all KA methods steadily improve as the overlap between the teacher's class sets increases. This is expected because these methods benefit from overlapping classes, particularly for instances where both teachers are experts as their logits will agree and the total predicted probability will increases. Finally, we find that our approach, TC, surpasses all other methods for the majority of settings by attaining higher accuracy, regardless of the % overlap. In some cases, other methods are competitive, but by and large TC is the dominant approach. Interestingly, for the HAR dataset, TC achieves its maximum performance with 0% overlap between the teachers' class sets. We hypothesize that this is a specific case where TC is able to pick



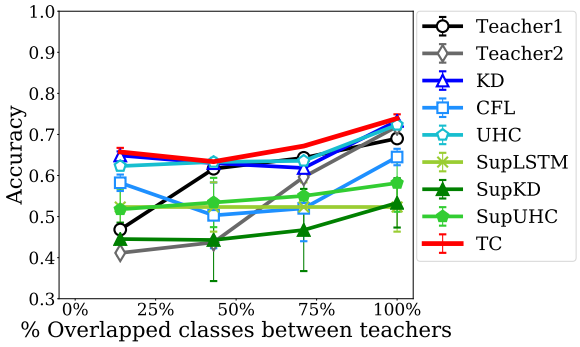
(a) SYN



(b) PED



(c) HAR



(d) ELEC

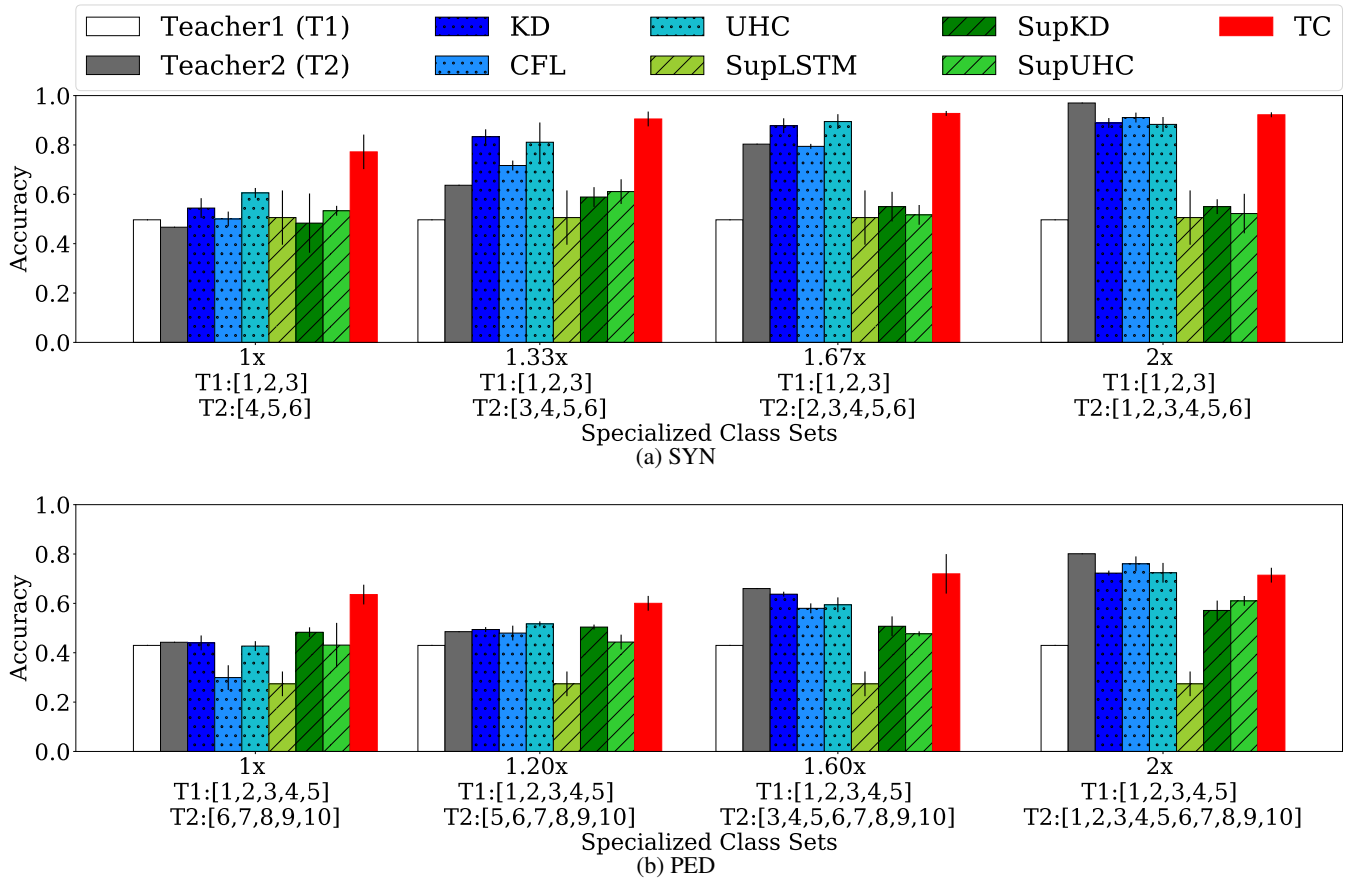
Figure 1: The expanding results presented in Figure 4 in our full paper. These performances are observed on SYN, PED, HAR, and ELEC over a wide variety of class-overlap ranges (0% to 100%).

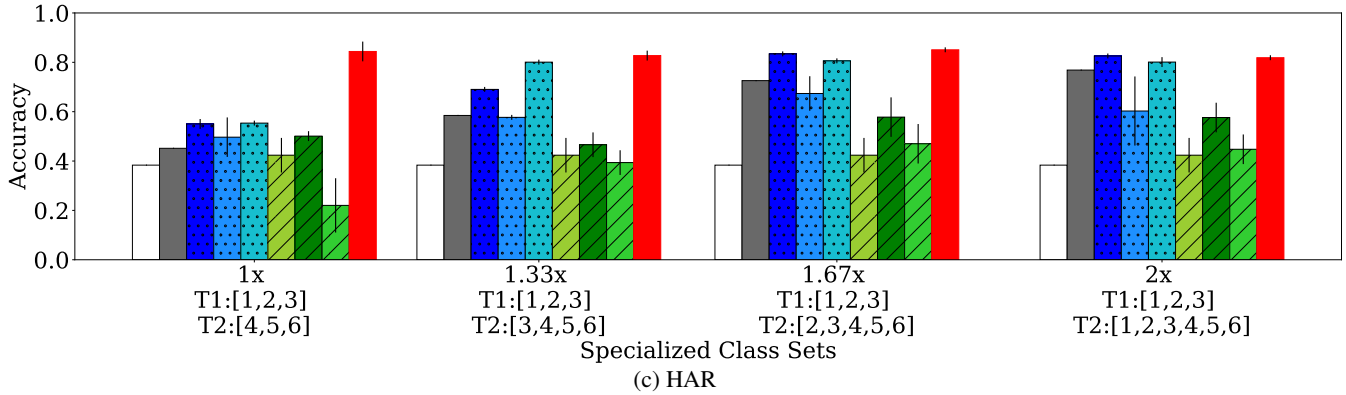
the correct teacher, even without overlap, because it does not rely on overlapping logits. Thus, as the % overlap in the teachers' class sets increases, there is no more improvement possible.

Overcoming overconfident teachers.

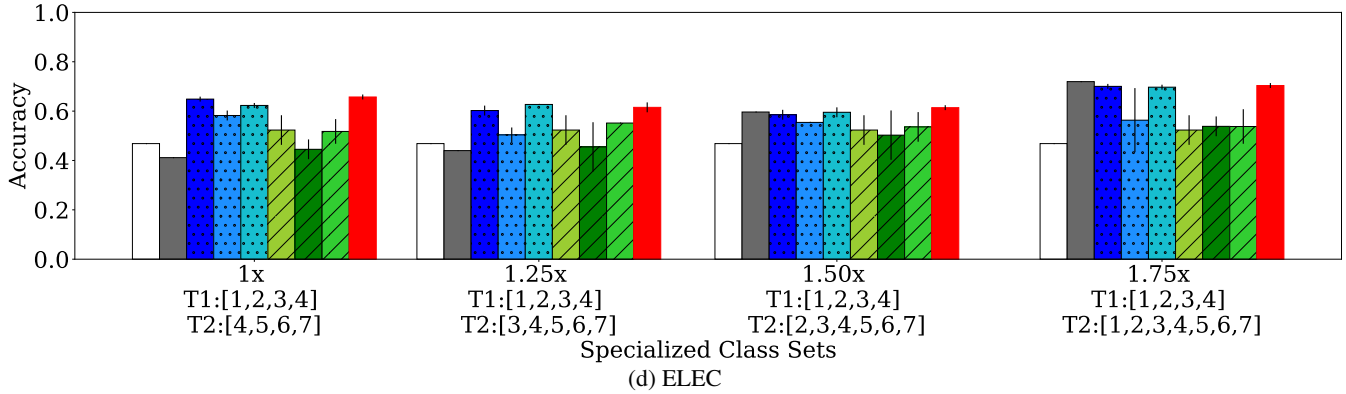
Here, we expand on our experiments regarding *overconfident* teachers (Figure 5 in the full paper), which make confident predictions for instances from classes they were not trained to predict. As described in the full paper, the sheer number of classes for which a teacher is an expert is a key contributor for overconfidence in predictions, which thereby dominate a different teacher's correct predictions. In all of the following experiments, we fix the number of classes for with Teacher 1 is an expert, and gradually increase the number of classes for which Teacher 2 is an expert. In our experiments, we stop when Teacher 2 covers all target classes which are twice as many classes as Teacher 1. We also assume again that only 2% of the training data are annotated. Figure 2 shows the comparison of each method with varying numbers of classes covered by each teacher. We first note that, again, the supervised methods (green bars) do not achieve competitive accuracy, as 2% labeling is not enough to train them effectively. Second, the KA methods (blue bars and our red bar) improve beyond both teachers, indicating they successfully learn to combine the knowledge of both teachers to some degree. However, our proposed method, TC (in red), surpasses the other KA methods to a very significant degree in most settings. Once Teacher 2 dominates Teacher 1 by enough (2x), all methods are roughly equivalent. In the balanced case (1x), TC is by far the winner in most cases. Even on the ELEC dataset, where the difference between compared methods is smaller, TC remains one of the strongest-performing methods for this task. The strong performance of TC in this experiment indicates that it is capable of picking the most-trustworthy teachers to rely on in the most settings.

Figure 2: The expanding results presented in Figure 5 in our full paper. These performances are observed on SYN, PED, HAR, and ELEC showing the compared results on different levels of overconfident teachers.




1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
1x
T1:[1,2,3]
T2:[4,5,6]
1.33x
T1:[1,2,3]
T2:[3,4,5,6]
1.67x
T1:[1,2,3]
T2:[2,3,4,5,6]
2x
T1:[1,2,3]
T2:[1,2,3,4,5,6]
Specialized Class Sets

(c) HAR


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
1x
T1:[1,2,3,4]
T2:[4,5,6,7]
1.25x
T1:[1,2,3,4]
T2:[3,4,5,6,7]
1.50x
T1:[1,2,3,4]
T2:[2,3,4,5,6,7]
1.75x
T1:[1,2,3,4]
T2:[1,2,3,4,5,6,7]
Specialized Class Sets

(d) ELEC