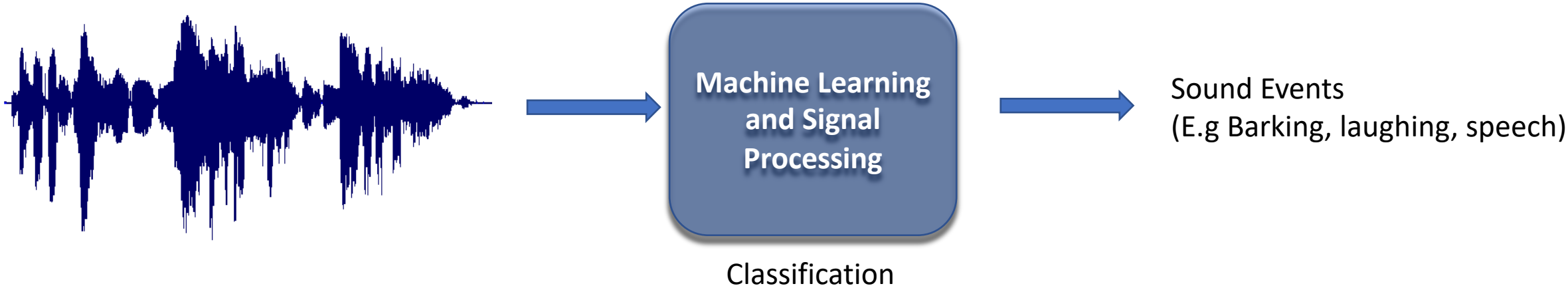
facebook
Reality Labs

ICASSP 2020

SeCoST: Sequential Co-Supervision For Large Scale Weakly Labeled Audio Event Detection

Anurag Kumar, Vamsi Krishna Ithapu

Audio Event Classification and Detection (AED)



Large scale learning of audio events

- Weakly labeled learning of sound events[1]
- Challenges with large scale weakly labeled learning of sounds
  - Large scale brings adverse learning conditions into the picture
  - Noise in labels is expected
  - Noise in audio data itself can become large

1. Kumar and Raj 2016

Large scale learning of audio events

- Learning a good model in one shot is hard

- Key Idea:

  - Human learning is not a one-shot process

  - Sequence of Teaching Shelves[2]

  - Learn Sequentially over multiple stages – Let previous stage(s) of learning guide future stages of learning

2. Minsky 1994.

SeCoST: Sequential Co-supervised Training

- Learn Sequentially over multiple stages
- Knowledge transfer through teacher-student framework[3]
- In SeCoST a cascade of neural networks are trained
- Network from prior stage guides training at current stage.

3. Hinton et. al. Distilling the knowledge in a neural network

SeCoST: Sequential Co-supervised Training

- How to guide future stages of learning ?

- Embed the knowledge of the model in the target space

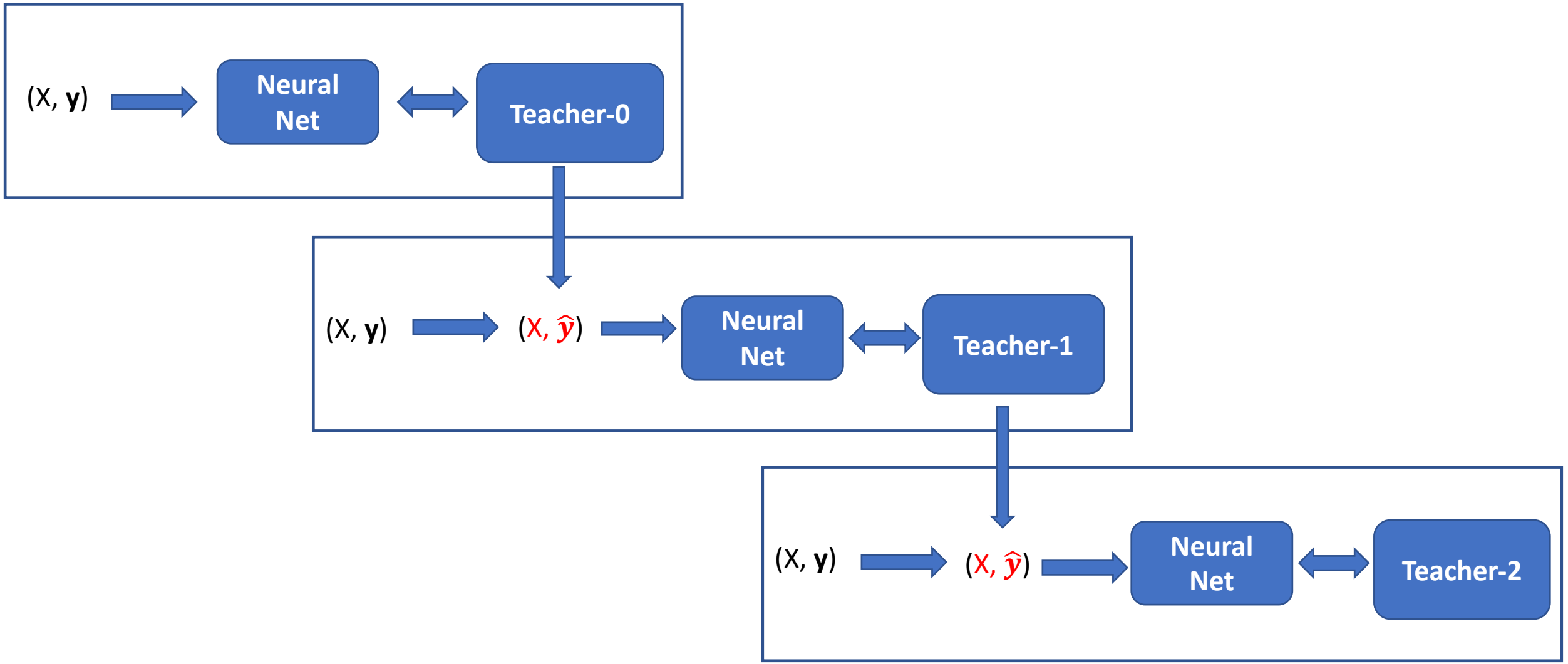
$$\bar{\mathbf{y}} = \alpha \mathbf{y} + (1 - \alpha)\hat{\mathbf{y}}$$

New target space

Original Labels

Prediction from current stage model

SeCoST: Sequential Co-supervised Training



SeCoST: Training losses

$$\mathcal{L}(\mathcal{N}(\theta, X), \bar{\mathbf{y}}) = \mathcal{L}(\mathcal{N}(\theta, X), \alpha\mathbf{y}) + (1-\alpha)\frac{1}{|C|}\sum_{i=1}^{|C|}\hat{y}_i log\frac{1-p_i}{p_i}$$

Supervision from Original Labels

Supervision from Teacher

$$\mathcal{L}(\mathcal{N}(\theta, X), \bar{\mathbf{y}}) = \mathcal{L}(\mathcal{N}(\theta, X), \mathbf{y}) + (1-\alpha)\frac{1}{|C|}\sum_{i=1}^{|C|}(y_i - \hat{y}_i) log\frac{1-p_i}{p_i}$$

1. Difference between teacher output and original in second term
2. Learning from how different teacher's predictions is from original labels

Experiments and Results

CNN Architecture

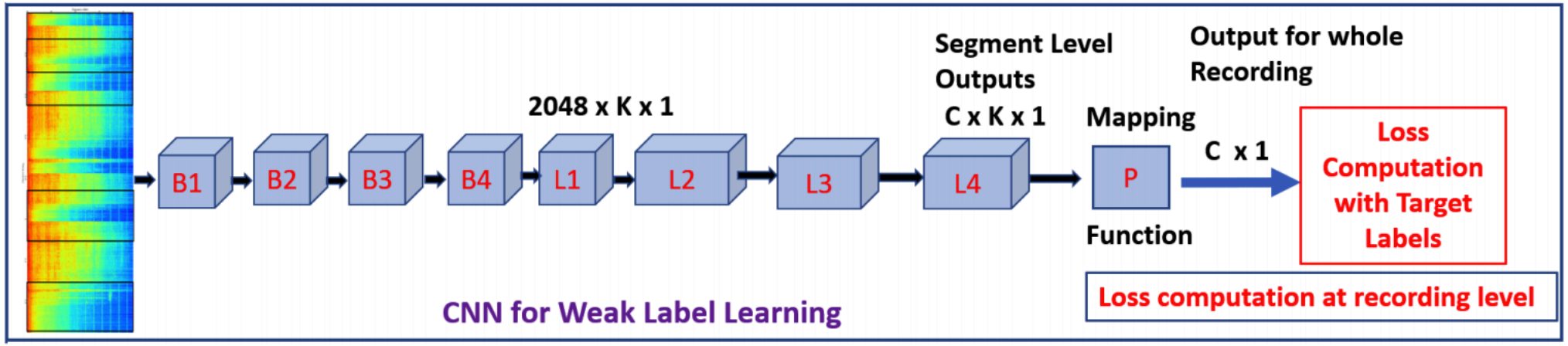


Fig. 1. **WELS-Net**: Deep CNN for weakly labeled AED. $|C|$: number of classes. K: number of segments obtained for the given input. P maps the segment level output(s) to the recording level output.

1. Can do both detection and recognition while learning from weakly labeled data
2. Details in paper.

Experiments

- Dataset: Audioset
  - 527 sound events
  - > 5000 hours of weakly labeled audio recordings
  - All results reported on Eval set of Audioset
    - Eval set contains around 20,000 10 second clips
  - Average precision for each class and mean average precision over all classes are used as evaluation metrics.

Experiments – Base Model (Teacher-0)

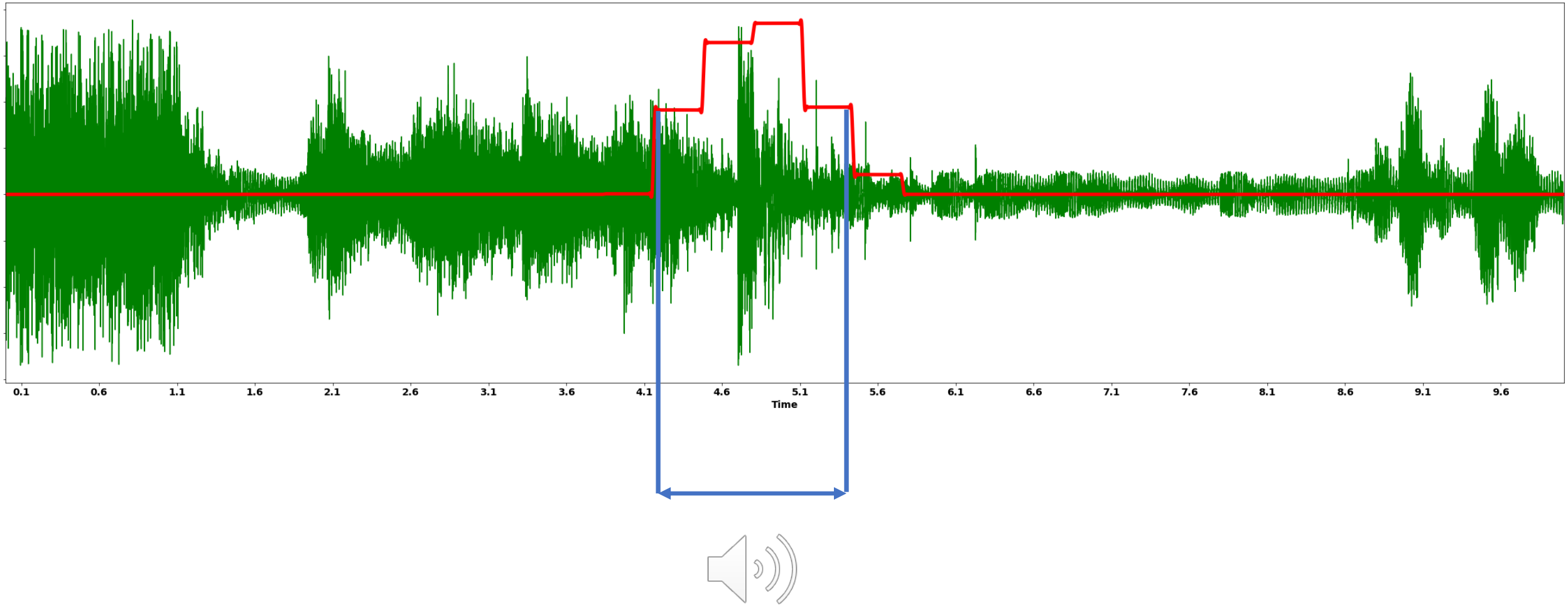
- Base model – trained on available labels

Method	Mean Average Precision
*Prior SOTA[4]	36.9
Our Base Model	35.2

*Uses features from a pre-trained network[5]. Ours uses logmel representations of audio recordings

4. Kong et. al. 2019 5. Hershey et. al. 2017

Experiments – Localization of events



Red Line – Score output for “Breaking” sound event

Experiments: SeCoST Performance Summary

Method	Mean Average Precision
*Prior SOTA	36.9
Our Base Model	35.2
SeCoST	**38.3**

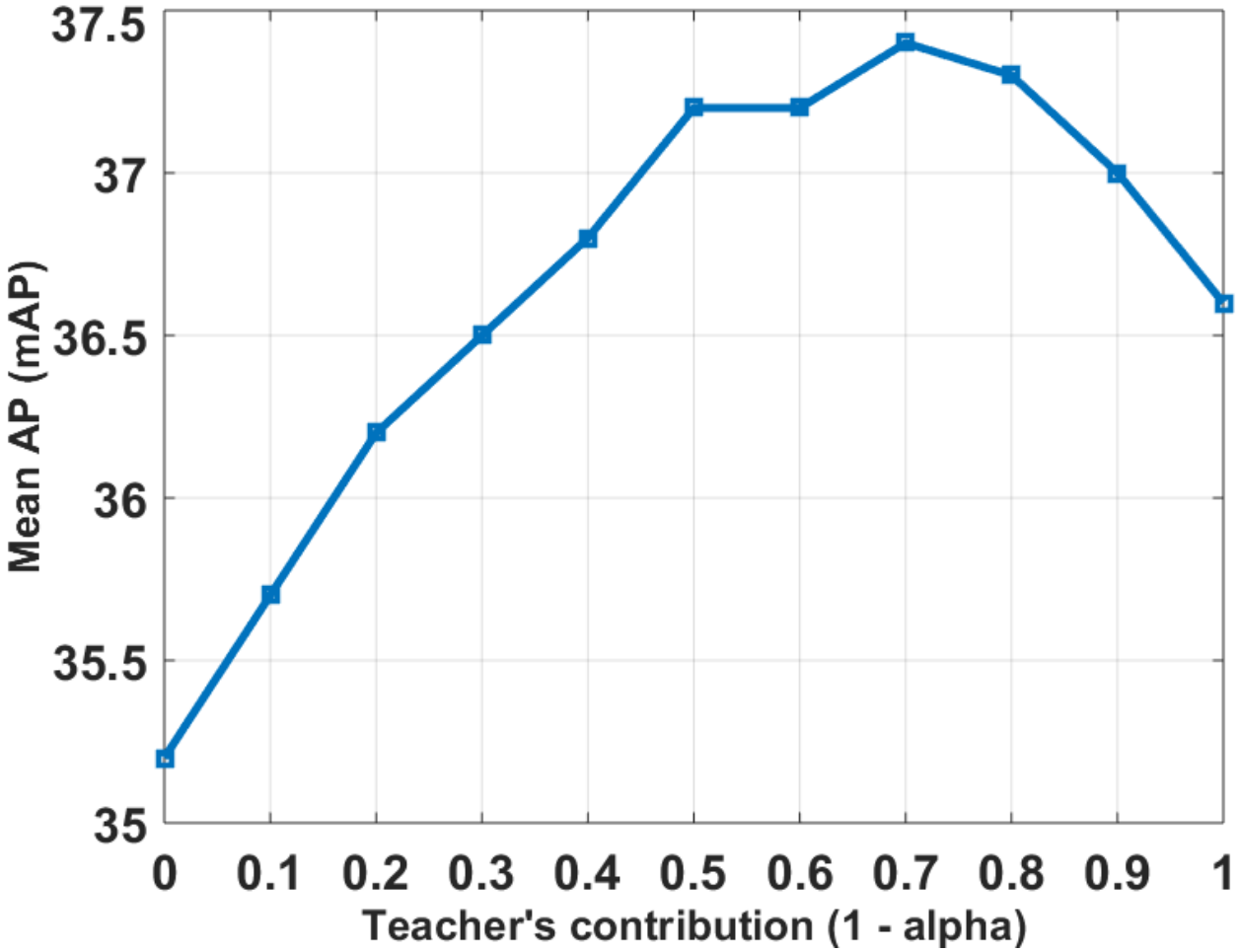
*** SeCoST improves the base model by ~8.8%**

SeCoST: Effect of α on performance

How does the weight teacher's contribution affect performance ?

x-axis shows the weight of teacher's contribution (1 – alpha) and y-axis shows

SeCoST for just 1 stage

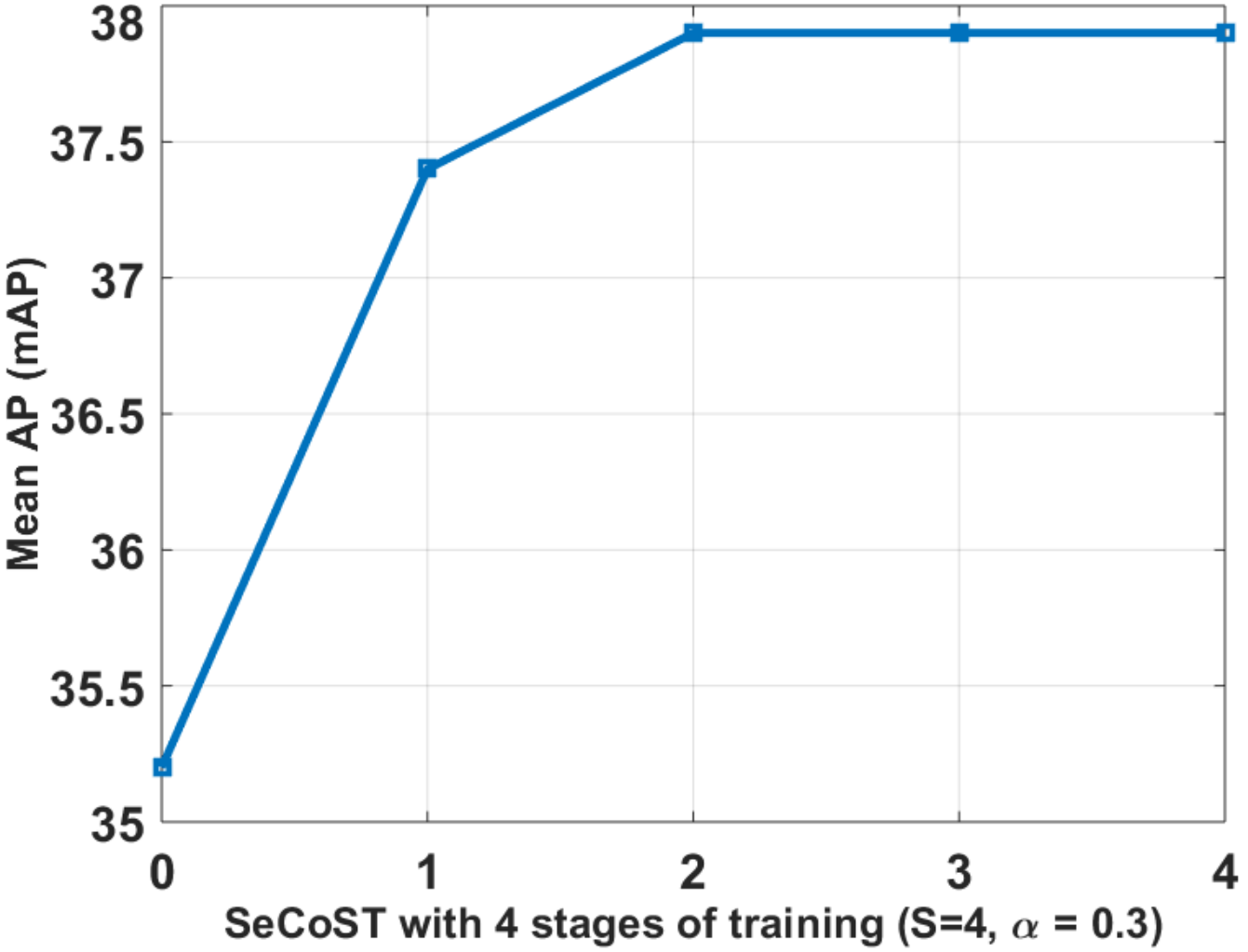


SeCoST: Multiple Stages with Fixed α

SeCoST applied for 4 stages.

Teacher at each stage is student from prior stage.

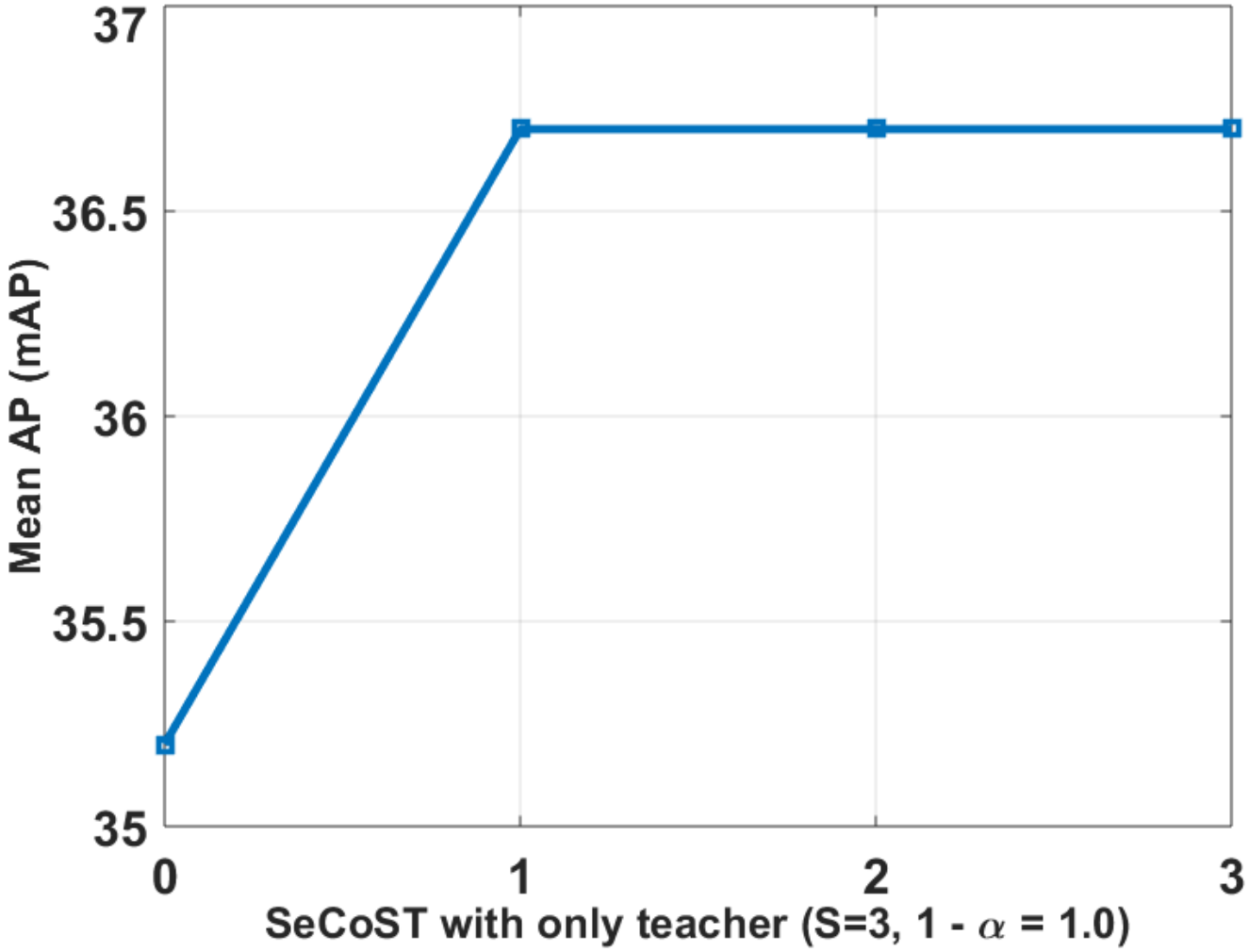
Teacher's contribution remains fixed (1-α = 0.7)



SeCoST: Multiple Stages with Fixed α

SeCoST applied for 3 stages.

SeCoST with full supervision from teacher only. Available labels are not used in SeCoSt (1-α = 1)

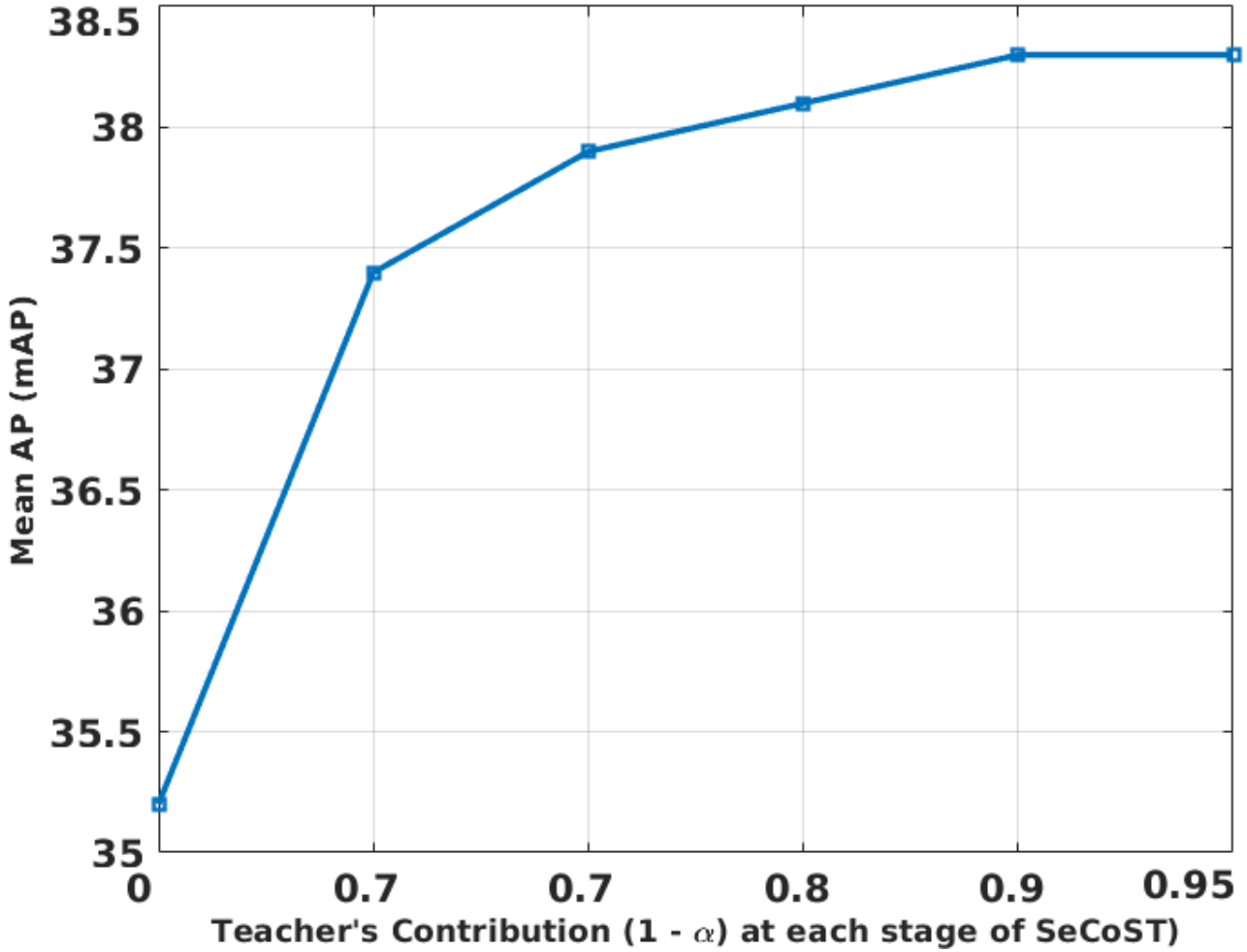


SeCoST: Multiple Stages with Increasing α

SeCoST applied for 5 stages.

Increase contribution as learning improves.

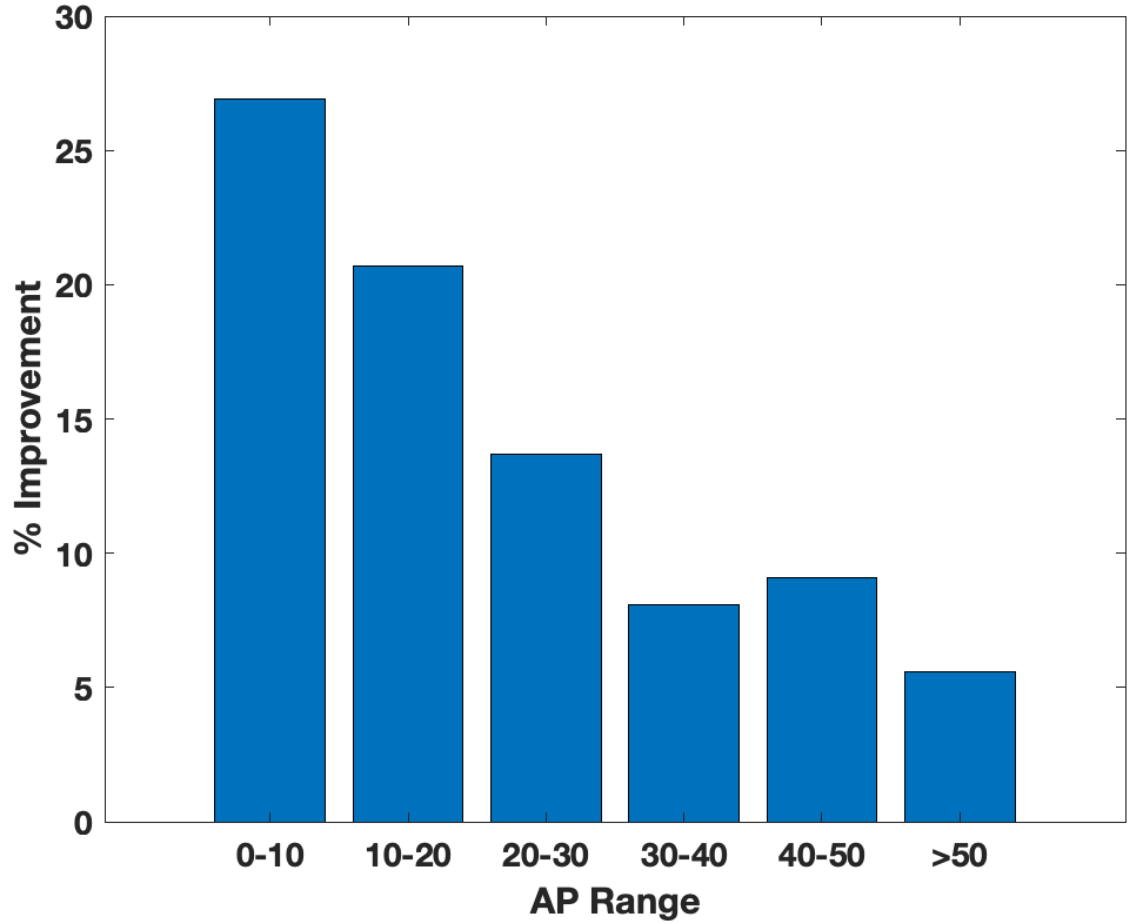
Teacher's contribution increases from 0.7 to 0.95



Some Class Specific Analysis

- Improvement for over 85% of classes
  - 448 out of 527

- Average improvement for classes in different ranges.



Thank You!