Reptile goes deeper deeper

Dongze Lian

Rethinking cross-modal problem

- Dongze Lian

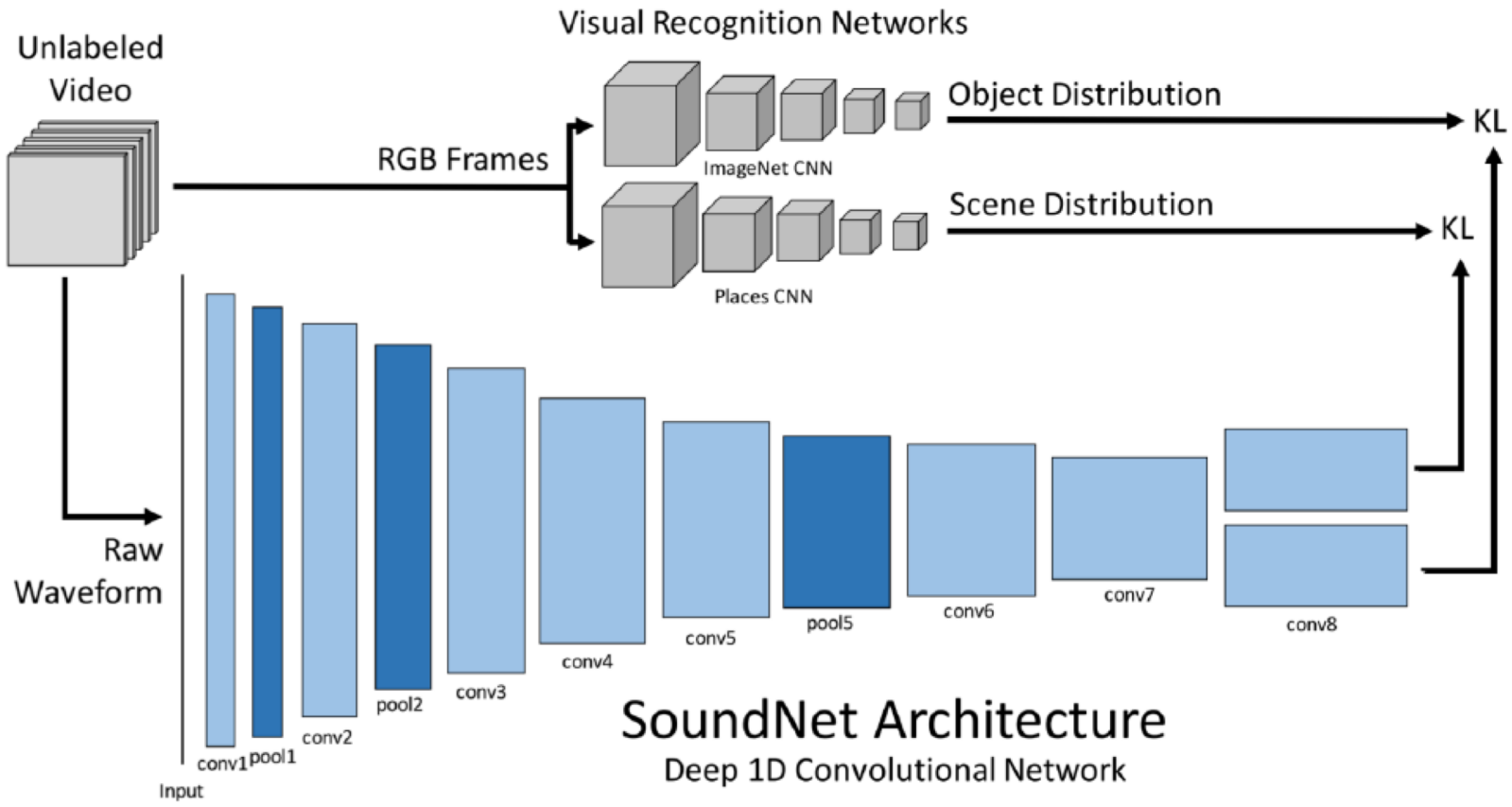
Outline

1. SoundNet http://soundnet.csail.mit.edu/
2. The Sound of Pixels http://sound-of-pixels.csail.mit.edu/
3. Radio-based pose estimation http://rfpose.csail.mit.edu/
4. Conclusion

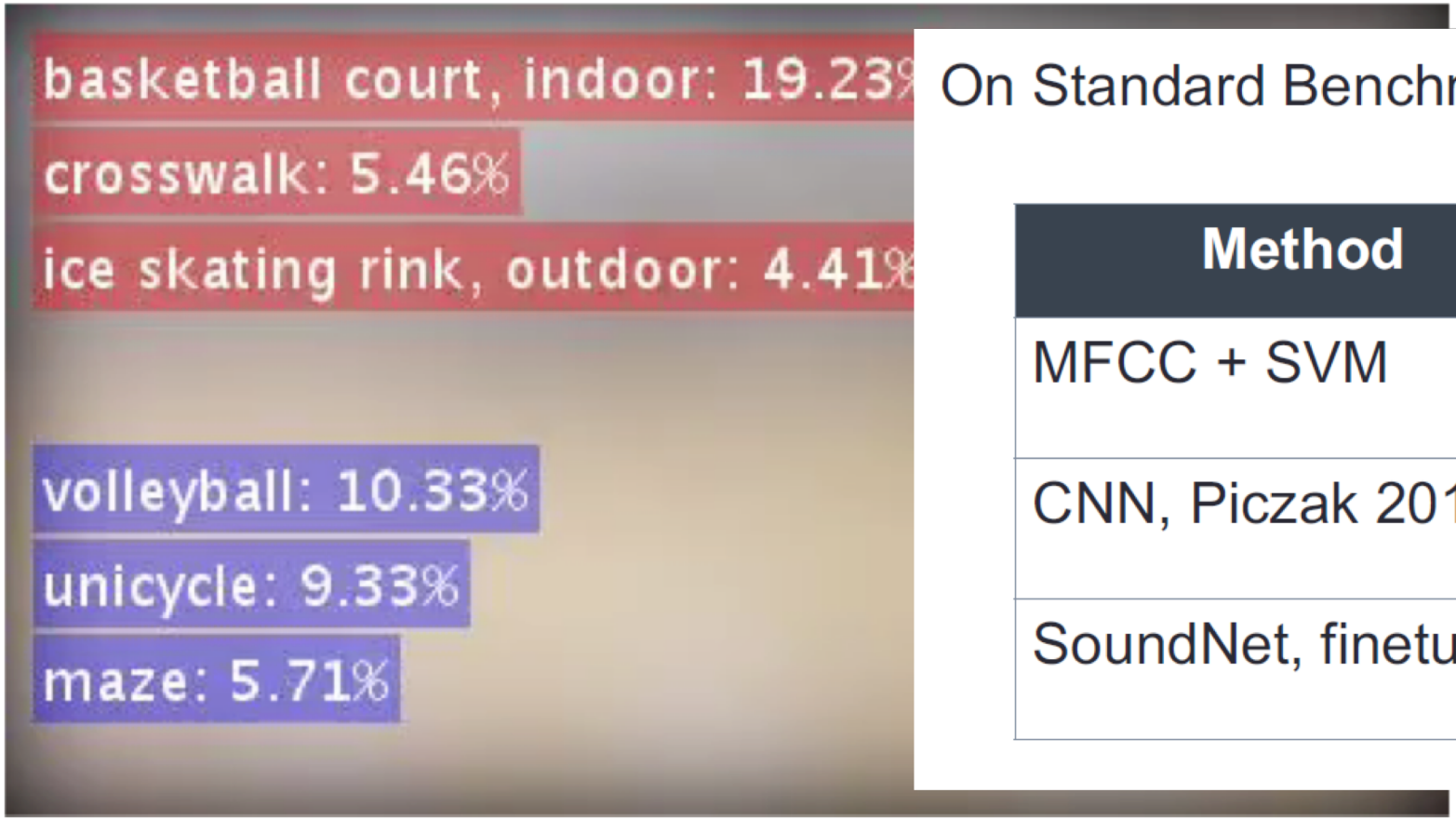
SoundNet Review

- We do not have an ImageNet for sound…

SoundNet Review



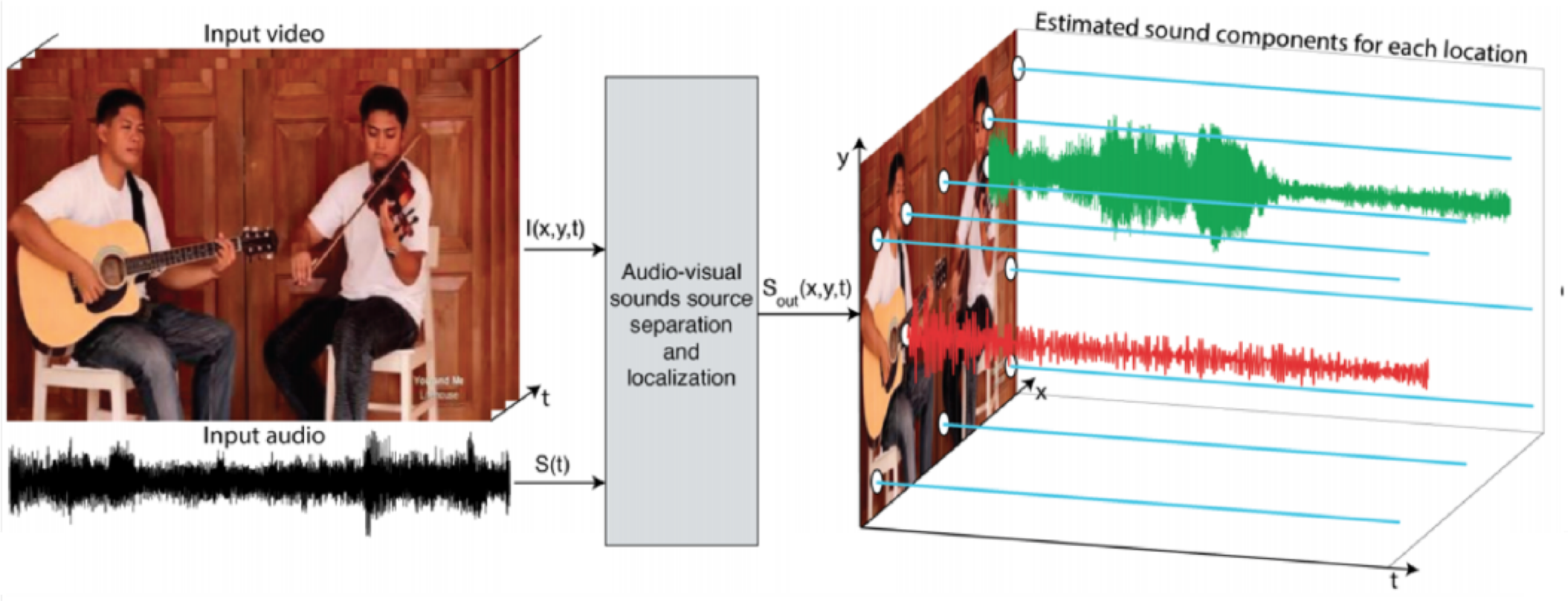
SoundNet Review



On Standard Benchmark ESC-50

Method	Accuracy
MFCC + SVM	39%
CNN, Piczak 2015	64%
SoundNet, finetune	74%

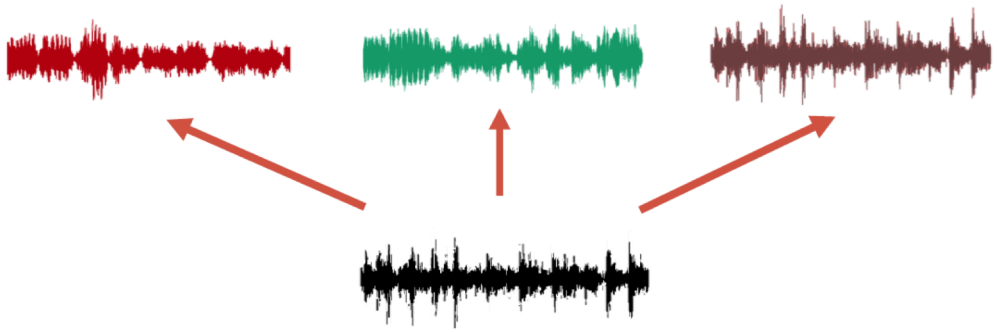
The Sound of Pixels Review



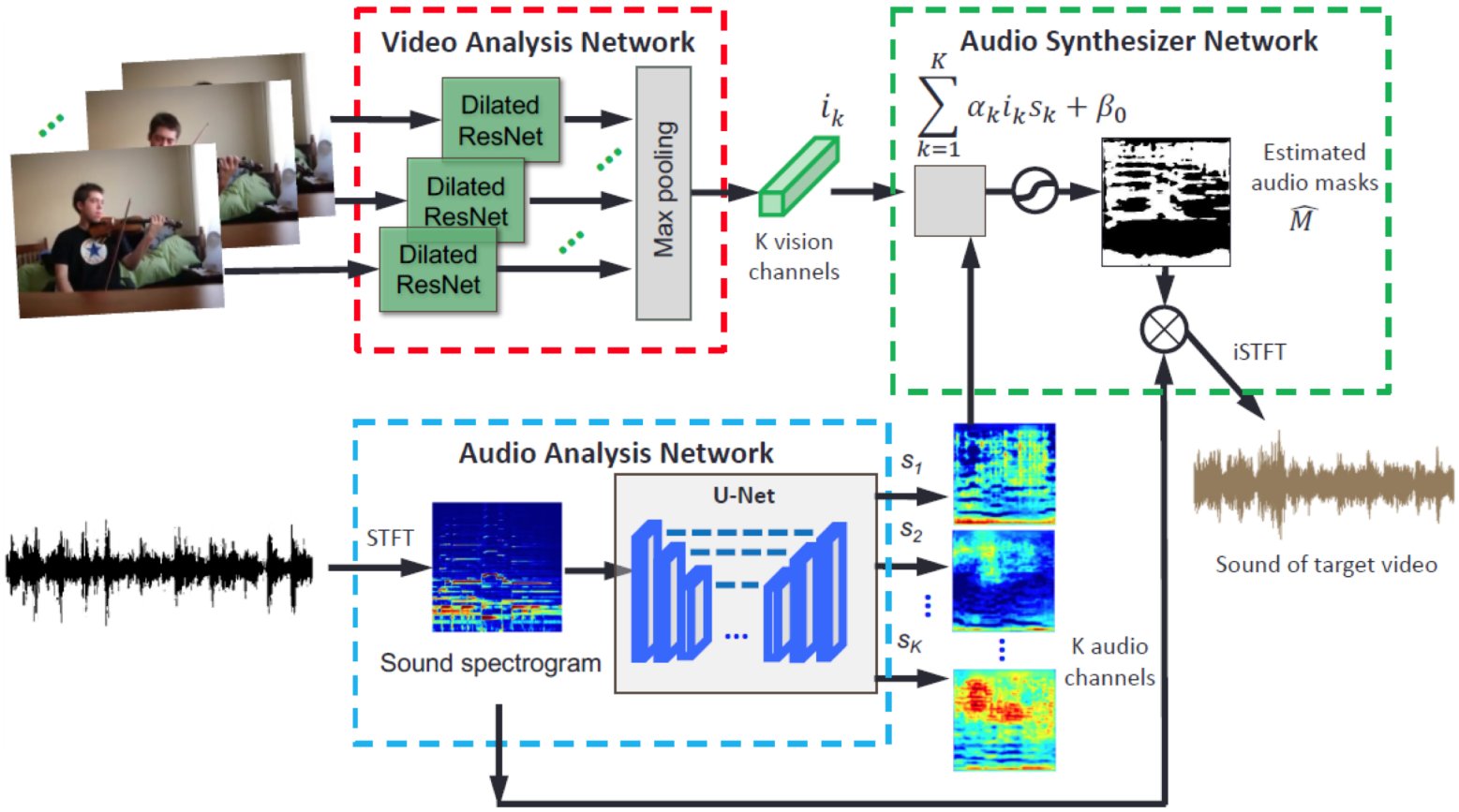
The Sound of Pixels Review

Audio-only:

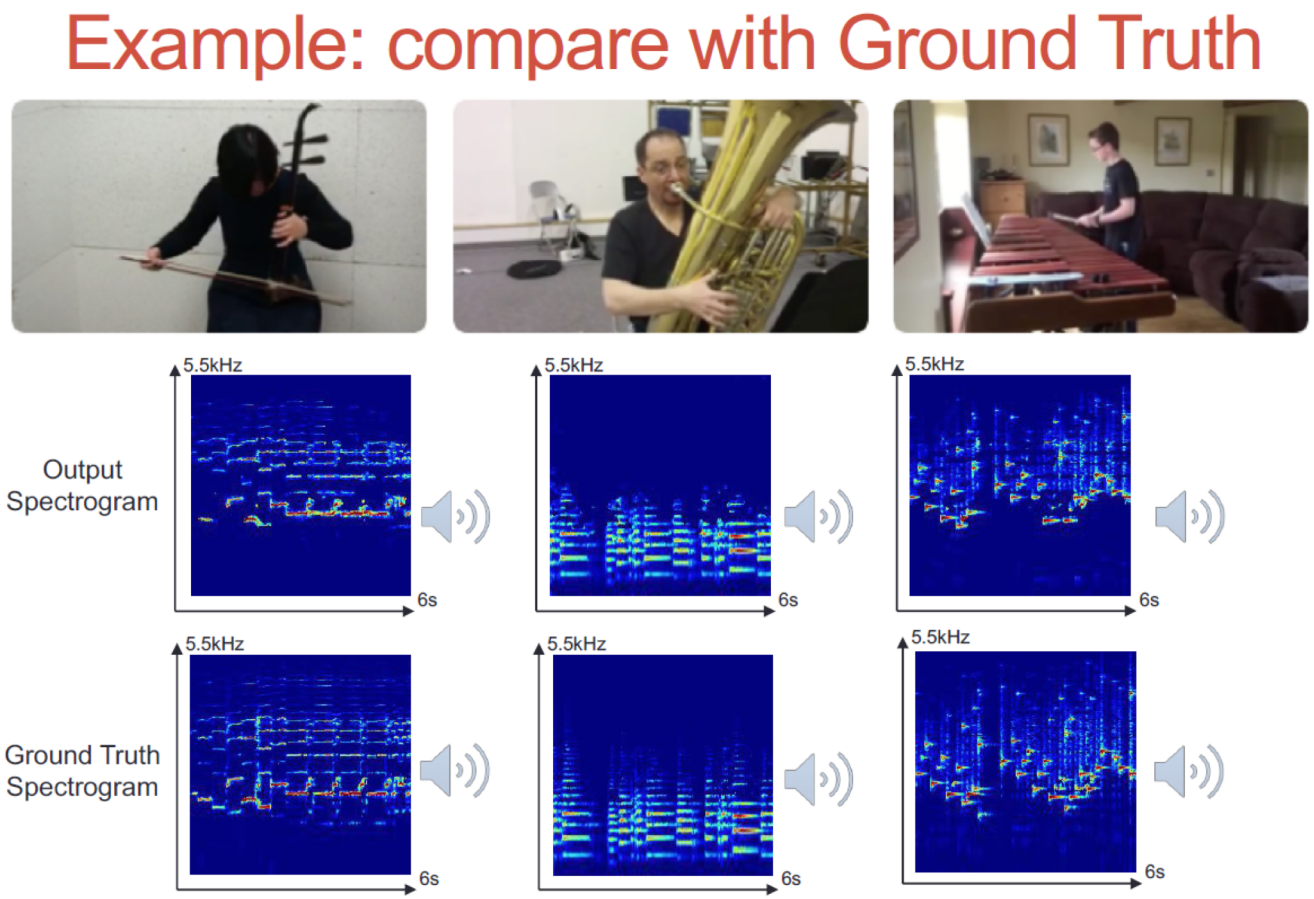
- Ill-posed, challenging
- Permutation problem if unsupervised



The Sound of Pixels Review



The Sound of Pixels Review



Radio-based pose estimation

- 1. Pose estimation task: 2D pose estimation
- 2. Difficulty: occlusion
- 3. Solution in the past work: infer based on the visible body

(deformable, fully occlusion)

- 4. Solution in this paper: ratio-based
- 5. Idea: WIFI traverse walls and reflect off the human body

Method

1. Encoding RF (radio frequency) signal:

FMCW (Frequency Modulated Continuous Wave), antenna arrays

Horizontal and vertical heat maps

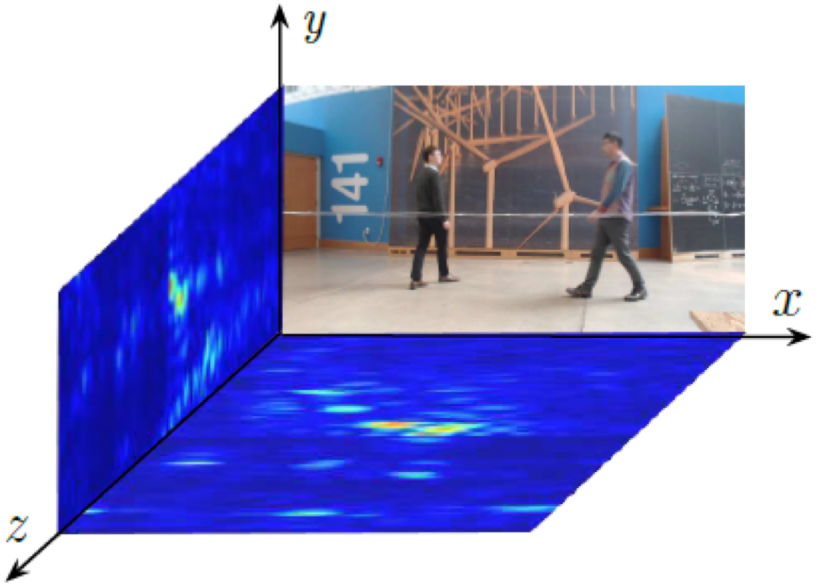


Figure 2: RF heatmaps and an RGB image recorded at the same time.

Method

2. Network

Teacher-student design

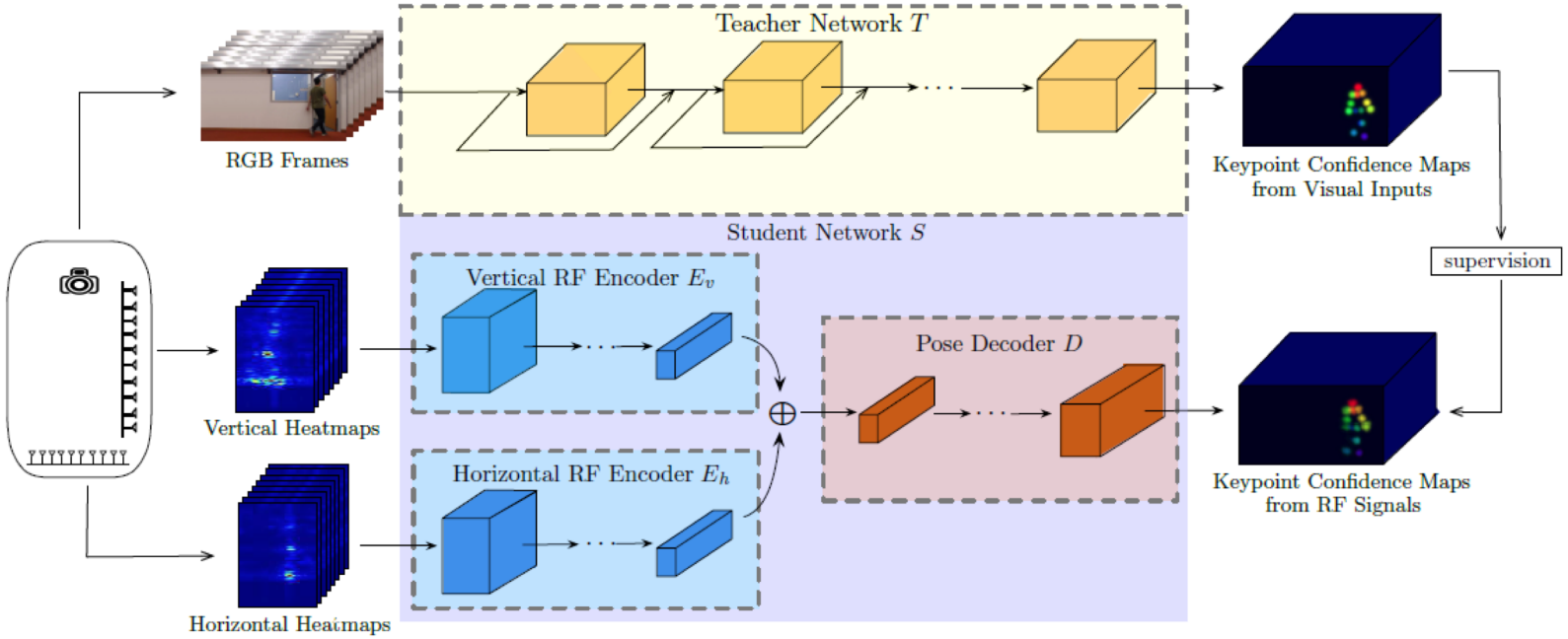


Figure 3: Our teacher-student network model used in RF-Pose. The upper pipeline provides training supervision, whereas the bottom pipeline learns to extract human pose using only RF heatmaps.

Dataset

- Visible scenes (70% for train, 30% for test)
- Through-walls (evaluate
- Ground-truth for evalua
  through-walls)

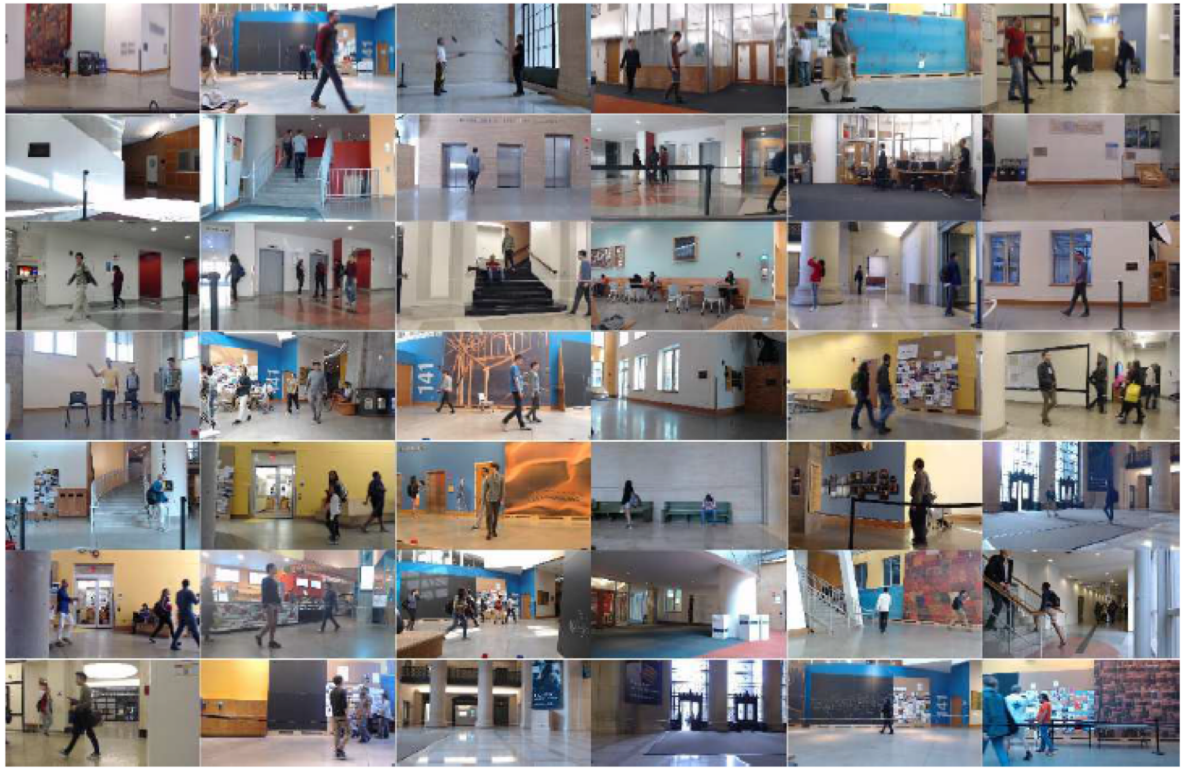


Figure 4: Different environments in the dataset.

Experiments

Methods	Visible scenes			Through-walls		
	AP	AP^{50}	AP^{75}	**AP**	AP^{50}	AP^{75}
RF-Pose	62.4	**93.3**	70.7	**58.1**	**85.0**	**66.1**
OpenPose[10]	**68.8**	77.8	**72.6**	-	-	-

Table 1: Average precision in visible and through-wall scenarios.

Conclusion

Pros: clever, different from multi-modal fusion

Consider:

- 1. How to supervise
- 2. Need to align or synchronize
- 3. A lot of unlabeled samples

Cons:

Supervised signal is generated by network. Align? Not sure. A lot of unlabeled scenes.